



UNIVERSITY OF  
LIVERPOOL

# **Biomarker-Guided Clinical Trial Designs**

---

Thesis submitted in accordance with the requirements

of the University of Liverpool for the degree of Doctor

in Philosophy

by

Miranta Antoniou

March 2018

# Acknowledgements

I wish to express my sincere appreciation to those who have contributed to this thesis and supported me during this doctoral journey.

First of all, I would like to give special thanks to my supervisors Dr Ruwanthi Kolamunnage Dona and Dr Andrea Jorgensen for their continuous guidance, advice and encouragement these past three years. I could not have imagined having better advisors for my PhD study.

I am also grateful to the Institute of Translational Medicine, University of Liverpool and the MRC North West Hub For Trials Methodology Research for funding my studies and to PhD students and staff of the Department of Biostatistics who have supported me throughout.

I would also like to take the opportunity to thank the Information Systems (IS) specialists working within the University of Liverpool Clinical Trials Research Centre (CTRC), namely Mr Andrew Ovens and Dr Duncan Appelbe for their support on developing an online tool for designing biomarker-guided clinical trials (BiGTed) which is presented in Chapter 4 of this thesis.

In addition, I thank the participants of the workshop held at the University of Liverpool in London Campus on the 15<sup>th</sup> March 2017, organized by the MRC Hubs for Trials Methodology Research Network's Stratified Medicine Working Group (SMWG) for sharing their experience related to the challenges of biomarker-guided trial designs in real practice. Chapter 8 of this thesis is based on the issues discussed at this workshop.

I would like to thank my PhD examiners, Dr Susanna Dodd and Dr James Wason for their valuable suggestions and comments which helped me to improve this thesis.

Special thanks go to my life-long friends Lina and Katerina for always listening and for their constant encouragement which helped me to see my doctoral study through to the end.

My heartfelt thanks go to my life companion Fotis, who has been a constant source of strength and inspiration. I thank him for his precious support throughout this experience and for his insightful discussions and suggestions.

Finally, I would especially like to thank all my family for their constant emotional support and for helping me survive all the stress and not letting me give up over these years.

# Abstract

**Title:** Biomarker-Guided Clinical Trial Designs

**Author:** Miranta Antoniou

Personalized medicine is a rapidly growing area of research which has attracted much attention in recent years in the field of medicine. The ultimate aim of this approach is to ensure that the most appropriate treatment which provides clinical benefit will be tailored to each patient according their personal characteristics. However, testing the effectiveness of a biomarker-guided approach to treatment in improving patient health yields challenges both in terms of trial design and analysis. Although a variety of biomarker-guided designs have been proposed recently, their statistical validity, application and interpretation has not yet been fully explored.

A comprehensive literature review based on an in-depth search strategy has been conducted with a view to providing researchers with clarity in definition, methodology and terminology of the various reported biomarker-guided trial designs. Additionally, a user-friendly online tool ([www.BiGTeD.org](http://www.BiGTeD.org)) informed by our review has been developed to help investigators embarking on such trials decide on the most appropriate design.

Simulation studies for the investigation of key statistical aspects of such trial designs and statistical approaches such as the sample size requirement under different settings have been performed.

Furthermore, a strategy has been applied to choose the most optimal design in a given setting where a previously proposed clinical trial proved inefficient due to the very large sample size that was required. Statistical techniques to calculate the corresponding sample size have been applied and an adaptive version of the proposed design has been explored through simulations.

Practical challenges of biomarker-guided trials in terms of funding, ethical and regulatory issues, recruitment, monitoring, statistical analysis plan, biomarker assessment and data sharing issues are also addressed in this thesis.

The different biomarker-guided designs proposed so far need to be better understood by the research community in terms of analysis and planning and practical application as their proper use and choice can increase the probability of success of clinical trials which will result in development of personalised treatments in the future. Therefore, with this PhD thesis, we contribute to the knowledge enhancement of researchers regarding these studies by providing essential information and presenting statistical issues arising in their implementation. We hope that this work will help scientists to choose the right clinical trial design in the era of personalized medicine which is of utmost importance for the translation of drug development into the improvement of human health.



## List of Tables

<b>Table 2.1.</b> Characteristics of biomarker-guided adaptive trial designs in Phase II and Phase III .....	13
<b>Table 3.1.</b> Types of Biomarker guided non-adaptive designs proposed within the last ten years.....	69
<b>Table 3.2.</b> Sample size formulae for biomarker-guided clinical trial designs.....	88
<b>Table 5.1.</b> Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 1 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the first scenario of hazard ratios and significance levels are also presented. ....	209
<b>Table 6.1.</b> Sample size of the Biomarker-strategy design with treatment randomisation in the control arm in each effect size scenario.....	231
<b>Table 6.2.</b> Sample size of the Marker Stratified design based on the target effect size and acamprosate response rate in each biomarker-defined subgroup. ....	242
<b>Table 6.3.</b> Sample size of the Sequential Subgroup-Specific design based on the target effect size and acamprosate response rate in biomarker-positive subgroup. ....	247
<b>Table 6.4.</b> Sample size of the Parallel Subgroup-Specific design based on the target effect size and acamprosate response rate in each biomarker-defined subgroup. ...	251
<b>Table 6.5.</b> Sample size of the Reverse Marker-Based strategy design with treatment randomization in the control arm in each effect size scenario. ....	256
<b>Table 6.6.</b> Required total number of patients for four potential designs applied to STRONG trial.....	258

<b>Table 6.7.</b> Required total number of events and corresponding hazard ratio of the Reverse Marker-Based strategy design applied to STRONG trial. ....	264
<b>Table 7.1.</b> Summary of simulation parameters for both binary and time-to-event outcomes. ....	282
<b>Table A.1.</b> Characteristics of variations of Biomarker-guided adaptive trial designs .....	371
<b>Table C.1.</b> Accrual rate and number of events and patients (calculated from (5.7), (5.1) and (5.3) respectively) which achieve approximate 80% power for different scenarios of hazard ratios and significance levels, and the corresponding power of each biomarker-defined subgroup yielded from the simulation. ....	386
<b>Table C.2.</b> Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 2 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the second scenario of hazard ratios and significance levels are also presented. ....	390
<b>Table C.3.</b> Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 3 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the third scenario of hazard ratios and significance levels are also presented. ....	393
<b>Table C.4.</b> Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 4 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from	

(5.1) and (5.3) respectively) which achieve 80% power for the fourth scenario of hazard ratios and significance levels are also presented.....396

**Table D.1.1.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method and O'Brien-Fleming decision boundaries. ....419

**Table D.1.2.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method and O'Brien-Fleming decision boundaries. ....420

**Table D.1.3.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method and Pocock decision boundaries. ....421

**Table D.1.4.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method and Pocock decision boundaries. ....422

**Table D.1.5.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method and O'Brien-Fleming decision boundaries.....423

**Table D.1.6.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method and O'Brien-Fleming decision boundaries. ....425

**Table D.1.7.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with effect-size ratio method and Pocock decision boundaries. ....428

**Table D.1.8.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method and Pocock decision boundaries. ....430

<b>Table D.2.1.</b> Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary. ....	433
<b>Table D.2.2.</b> Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary. ....	434
<b>Table D.2.3.</b> Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method, Pocock efficacy boundaries and Pocock futility boundary. ....	435
<b>Table D.2.4.</b> Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method, Pocock efficacy boundaries and Pocock futility boundary. ....	436
<b>Table D.2.5.</b> Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary. ....	437
<b>Table D.2.6.</b> Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary. ....	439
<b>Table D.2.7.</b> Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method, Pocock efficacy boundaries and Pocock futility boundary. ....	442
<b>Table D.2.8.</b> Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method, Pocock efficacy boundaries and Pocock futility boundary. ....	444

## List of Figures

<b>Figure 2.1.</b> CONSORT diagram of the review process.....	9
<b>Figure 2.2.</b> Adaptive signature design. “R” refers to randomization of patients. ....	23
<b>Figure 2.3.</b> Outcome-based adaptive randomization design. “R” refers to randomization of patients.....	26
<b>Figure 2.4.</b> Outcome-based adaptive randomization design. “R” refers to randomization of patients.....	28
<b>Figure 2.5.</b> Adaptive patient enrichment design. “R” refers to randomization of patients. ....	31
<b>Figure 2.6.</b> Adaptive parallel Simon two-stage design. “R” refers to randomization of patients. ....	33
<b>Figure 2.7.</b> Multi-arm multi-stage (MAMS) design. “R” refers to randomization of patients. ....	37
<b>Figure 2.8.</b> Stratified adaptive design. “R” refers to randomization of patients.....	39
<b>Figure 2.9.</b> Tandem two stage design. “R” refers to randomization of patients.....	42
<b>Figure 3.1.</b> Flow diagram of the review process. From our search strategy a total number of 211 papers have been identified giving information regarding not only the biomarker-guided designs but also general information about personalized medicine and biomarkers. Before arriving at 211 papers, books, web pages for actual trials and papers published before 2005 were excluded. The 211 papers are split into two overlapping sets of 100 and 107 papers. The total of 207 is less than 211 due to overlap of papers, and also due to the fact that some articles referring to general information about personalized medicine and biomarkers and articles which do not provide further information on each broad of biomarker-guided designs were excluded. The	

107 papers for biomarker-guided adaptive trial designs were reviewed in our published paper Antoniou et al. (2016) [35].	67
<b>Figure 3.2.</b> Single arm designs	105
<b>Figure 3.3.</b> Enrichment designs. “R” refers to randomization of patients.	107
<b>Figure 3.4.</b> Marker Stratified designs. “R” refers to randomization of patients. ....	115
<b>Figure 3.5.</b> Sequential Subgroup-Specific design. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. ....	122
<b>Figure 3.6.</b> Parallel Subgroup-Specific design. “R” refers to randomization of patients. ....	125
<b>Figure 3.7.</b> Biomarker-positive and overall strategies with parallel assessment. “R” refers to randomization of patients. ....	127
<b>Figure 3.8.</b> Biomarker-positive and overall strategies with sequential assessment. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. ....	129
<b>Figure 3.9.</b> Biomarker-positive and overall strategies with fall-back analysis. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. ....	131
<b>Figure 3.10.</b> Marker Sequential test design (MaST). “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. ....	133
<b>Figure 3.11.</b> Hybrid design. “R” refers to randomization of patients. ....	136
<b>Figure 3.12.</b> Biomarker-strategy design with biomarker assessment in the control arm. “R” refers to randomization of patients.	139

<b>Figure 3.13.</b> Biomarker-strategy design without biomarker assessment in the control arm. “R” refers to randomization of patients.....	142
<b>Figure 3.14.</b> Biomarker-strategy design with treatment randomization in the control arm. “R” refers to randomization of patients.....	145
<b>Figure 3.15.</b> Reverse Marker-Based strategy design. “R” refers to randomization of patients. ....	147
<b>Figure 3.16.</b> Randomized Phase II trial design with biomarkers. “R” refers to randomization of patients. CI refers to the confidence interval. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. ....	150
<b>Figure 4.1.</b> ‘Pop-up’ box illustration.....	172
<b>Figure 4.2.</b> Online tool’s homepage.....	173
<b>Figure 4.3.</b> Example of the webpage of a distinct adaptive design .....	174
<b>Figure 4.4.</b> Example of an expanded version of an adaptive trial graphic .....	175
<b>Figure 4.5.</b> Example of an expanded version of an adaptive design graphic with the ‘pop-up’ box showing further information .....	175
<b>Figure 4.6.</b> Methodology information in the ‘Details’ section of an adaptive design graphic. ....	176
<b>Figure 4.7.</b> Statistical and practical information in the ‘Details’ section of an adaptive design graphic. ....	176
<b>Figure 4.8.</b> Key references in the ‘Details’ section of an adaptive design graphic...	177
<b>Figure 4.9.</b> ‘Variations’ section of an adaptive design graphic.....	177

<b>Figure 4.10.</b> Example of the webpage of a distinct non-adaptive design .....	178
<b>Figure 4.11.</b> Example of an expanded version of a shrunken non-adaptive design graphic.....	179
<b>Figure 4.12.</b> Example of an expanded non-adaptive design graphic with a ‘pop-up’ box showing further information .....	179
<b>Figure 4.13.</b> Utility information in the ‘Details’ section of a non-adaptive graphic	180
<b>Figure 4.14.</b> Methodology information in the ‘Details’ section of a non-adaptive graphic.....	180
<b>Figure 4.15.</b> Sample size formulae in the ‘Details’ section of a non-adaptive graphic .....	181
<b>Figure 4.16.</b> Statistical and practical information in the ‘Details’ section of a non-adaptive graphic.....	181
<b>Figure 4.17.</b> Key references in the ‘Details’ section of a non-adaptive graphic .....	182
<b>Figure 4.18.</b> Design-specific webpage of a non-adaptive design divided in different sub-categories. ....	183
<b>Figure 4.19.</b> Design-specific webpage of a sub-category of a non-adaptive design	183
<b>Figure 5.1.</b> Parallel Subgroup-Specific design. “R” refers to randomization of patients. a refers to the overall significance level between the two biomarker subgroup tests such that $a = a_- + a_+$ .....	188
<b>Figure 5.2.</b> A, B, C represent the required number of events and D, E, F represent the required number of patients of each biomarker-defined subgroup which achieve 80% power versus the corresponding hazard ratio for each of the three scenarios of significance levels. Figure 5.2 A and D corresponds to the significance levels $a_- = a_+ = 0.0125$ , Figure 5.2 B and E corresponds to the significance levels $a_- = 0.015$ and	



$\alpha_+ = 0.010$  and Figure 5.2 C and F corresponds to the significance levels  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$ ..... 196

**Figure 5.3.** A, B represent the required number of events and C, D represent the required number of patients which achieve 80% power versus the hazard ratio in each of the three scenarios of significance levels for each biomarker-defined subgroup separately. Figure 5.3 A and C corresponds to the biomarker-negative subgroup and the following significance levels: (i)  $\alpha_- = 0.0125$ , (ii)  $\alpha_- = 0.015$  and (iii)  $\alpha_- = 0.010$ . Figure 5.3 B and D corresponds to the biomarker-positive subgroup and the following significance levels: (i)  $\alpha_+ = 0.0125$ , (ii)  $\alpha_+ = 0.010$  and (iii)  $\alpha_+ = 0.015$ . ..... 198

**Figure 5.4.** Adaptive Parallel Subgroup-Specific design. “R” refers to randomization of patients.  $D_{1+}$  and  $D_{1-}$  correspond to the target number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively at the first stage of the study.  $D_+$  and  $D_-$  correspond to the total required number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively which are planned according to the non-adaptive approach.  $D_{2+}$  and  $D_{2-}$  correspond to the number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively at the second stage of the study..... 200

**Figure 5.5.** Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 1 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively..... 213

**Figure 5.6.** Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 25%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined

subgroup when (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$  respectively.....215

**Figure 6.1.** Biomarker-strategy design with treatment randomization in the control arm adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and Acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.228

**Figure 6.2.** Rate of return to any drinking for acamprosate .....230

**Figure 6.3.** Marker Stratified design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients. ....239

**Figure 6.4.** Sequential Subgroup Specific design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.....246

**Figure 6.5.** Parallel Subgroup-Specific design adopted in STRONG trial. “R” refers to randomisation of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients. ....249

**Figure 6.6.** Reverse Marker-Based strategy design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment and acamprosate corresponds to the control treatment. Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients..... 254

**Figure 7.1.** Type I error versus the information fraction of Simulation study 1. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025). ..... 287

**Figure 7.2.** Type I error versus the information fraction of Simulation study 2. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025). ..... 288

**Figure 7.3.** Power versus the information fraction under the alternative hypothesis of Simulation study 1. The horizontal line represents the power of the non-adaptive design (i.e. 80%). ..... 289

**Figure 7.4.** Power versus the information fraction under the alternative hypothesis of Simulation study 2. The horizontal line represents the power of the non-adaptive design (i.e. 80%). ..... 290

**Figure 7.5.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1. The horizontal line represents the sample size of the non-adaptive design..... 291

**Figure 7.6.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 2. The horizontal line represents the sample size of the non-adaptive design..... 291

**Figure 7.7.** Sample size versus the information fraction under the null hypothesis of Simulation study 1. The horizontal line represents the sample size of the non-adaptive design..... 292

<b>Figure 7.8.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 2. The horizontal line represents the sample size of the non-adaptive design.....	293
<b>Figure 7.9.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1.....	294
<b>Figure 7.10.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2.....	294
<b>Figure 7.11.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1. ....	295
<b>Figure 7.12.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2. ....	296
<b>Figure 7.13.</b> Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2.....	297
<b>Figure 7.14.</b> Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2. ....	298
<b>Figure 7.15.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025). ....	299
<b>Figure 7.16.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025). ....	299
<b>Figure 7.17.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025). ....	300

<b>Figure 7.18.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).....	300
<b>Figure 7.19.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).....	301
<b>Figure 7.20.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).....	301
<b>Figure 7.21.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).....	302
<b>Figure 7.22.</b> Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).....	302
<b>Figure 7.23.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.....	304
<b>Figure 7.24.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.....	305
<b>Figure 7.25.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.....	305

<b>Figure 7.26.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design. ....	306
<b>Figure 7.27.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio (i.e. 0.746). The horizontal line represents the power of the non-adaptive design. ....	307
<b>Figure 7.28.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio (i.e. 0.807). The horizontal line represents the power of the non-adaptive design. ....	307
<b>Figure 7.29.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio (i.e. 0.845). The horizontal line represents the power of the non-adaptive design. ....	308
<b>Figure 7.30.</b> Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the power of the non-adaptive design. ....	308
<b>Figure 7.31.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio (i.e. 0.746)..	309
<b>Figure 7.32.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio (i.e. 0.746)..	310
<b>Figure 7.33.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio (i.e. 0.845). The horizontal line represents the sample size of the non-adaptive design.....	310
<b>Figure 7.34.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio (i.e. 0.845). The horizontal line represents the sample size of the non-adaptive design.....	311

<b>Figure 7.35.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio (i.e. 0.807). The horizontal line represents the sample size of the non-adaptive design. ....	312
<b>Figure 7.36.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio (i.e. 0.807). The horizontal line represents the sample size of the non-adaptive design. ....	312
<b>Figure 7.37.</b> Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the sample size of the non-adaptive design.....	313
<b>Figure 7.38.</b> Sample size versus the information fraction under the alternative and null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the sample size of the non-adaptive design.	314
<b>Figure 7.39.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	315
<b>Figure 7.40.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	316
<b>Figure 7.41.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	316
<b>Figure 7.42.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	317

<b>Figure 7.43.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	318
<b>Figure 7.44.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	318
<b>Figure 7.45.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	319
<b>Figure 7.46.</b> Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design. ....	319
<b>Figure 7.47.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio.	320
<b>Figure 7.48.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio.	321
<b>Figure 7.49.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio. ....	321
<b>Figure 7.50.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio. ....	322
<b>Figure 7.51.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio. ....	322



<b>Figure 7.52.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio.	323
<b>Figure 7.53.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio.	323
<b>Figure 7.54.</b> Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio.	324
<b>Figure 7.55.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio.	325
<b>Figure 7.56.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio.	325
<b>Figure 7.57.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio.	326
<b>Figure 7.58.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio.	326
<b>Figure 7.59.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio.	327
<b>Figure 7.60.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio.	327
<b>Figure 7.61.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio.	328

<b>Figure 7.62.</b> Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. ....	328
<b>Figure 7.63.</b> Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio. ....	329
<b>Figure 7.64.</b> Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio. ....	330
<b>Figure 7.65.</b> Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio. ....	330
<b>Figure 7.66.</b> Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. ....	331
<b>Figure 7.67.</b> Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio. ....	332
<b>Figure 7.68.</b> Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio. ....	332
<b>Figure 7.69.</b> Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio. ....	333
<b>Figure 7.70.</b> Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. ....	333
<b>Figure B.1.</b> Sequential before–after pharmacogenetic diagnostic study .....	381
<b>Figure B.2.</b> Classifier randomization design. “R” refers to randomization of patients. ....	382

<b>Figure B.3.</b> Modified marker strategy design. “R” refers to randomization of patients.	383
--	-----

<b>Figure B.4.</b> Two-way Stratified design (for validation of prognostic biomarkers). “R” refers to randomization of patients.	385
--	-----

<b>Figure C.1.</b> Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 2 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i) $a_- = a_+ = 0.0125$ , (ii) $a_- = 0.015$ and $a_+ = 0.010$ and (iii) $a_- = 0.010$ and $a_+ = 0.015$ respectively.	399
---	-----

<b>Figure C.2.</b> Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 3 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i) $a_- = a_+ = 0.0125$ , (ii) $a_- = 0.015$ and $a_+ = 0.010$ and (iii) $a_- = 0.010$ and $a_+ = 0.015$ respectively.	400
---	-----

<b>Figure C.3.</b> Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 4 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i) $a_- = a_+ = 0.0125$ , (ii) $a_- = 0.015$ and $a_+ = 0.010$ and (iii) $a_- = 0.010$ and $a_+ = 0.015$ respectively.	401
---	-----

<b>Figure C.4.</b> Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 50%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined subgroup when (i) $a_- = a_+ = 0.0125$ , (ii) $a_- = 0.015$ and $a_+ = 0.010$ and (iii) $a_- = 0.010$	
---	--

and  $a_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$ ..... 402

**Figure C.5.** Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 75%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined subgroup when (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$ ..... 403

## List of Publications

1. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. PLoS ONE. 2016;11(2):e0149803. doi: 10.1371/journal.pone.0149803.
2. Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. Journal of Personalized Medicine. 2017;7(1). doi: 10.3390/jpm7010001.
3. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Fixed and adaptive Parallel Subgroup-Specific design for survival outcomes: power and sample size. 2017 (Under review)

# Contents

<b>Acknowledgements .....</b>	<b>i</b>
<b>Abstract .....</b>	<b>iii</b>
<b>List of Tables.....</b>	<b>iv</b>
<b>List of Figures .....</b>	<b>viii</b>
<b>List of Publications .....</b>	<b>xxiv</b>
<b>Contents .....</b>	<b>xxv</b>
<b>Chapter 1. Introduction.....</b>	<b>1</b>
1.1. Personalized medicine.....	1
1.1.1. Biomarker-guided clinical trial designs .....	2
1.2. Novel insight into Biomarker-guided clinical trial designs .....	2
1.2.1. Creating a user-friendly, interactive web resource .....	3
1.2.2. Choosing the best trial design .....	3
1.2.3. Guidance on practical challenges .....	4
1.3. References .....	4
<b>Chapter 2. Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review .....</b>	<b>6</b>
2.1. Abstract.....	<b>Error! Bookmark not defined.</b>
2.2. Introduction .....	6
2.3. Methods and Findings.....	8
2.3.1. Adaptive designs.....	10
2.3.1.1. Definitions and terminology .....	10
2.3.1.2. Adaptations to the design.....	11
2.3.1.3. Analysis of adaptive designs.....	11
2.3.2. Biomarker-guided adaptive trial designs.....	12

2.3.2.1. Adaptive signature design.....	22
2.3.2.2. Outcome-based adaptive randomization design .....	24
2.3.2.3. Adaptive threshold sample-enrichment design .....	27
2.3.2.4. Adaptive patient enrichment design.....	30
2.3.2.5. Adaptive parallel Simon two-stage design .....	32
2.3.2.6. Multi-arm multi-stage designs.....	36
2.3.2.7. Stratified adaptive design .....	39
2.3.2.8. Tandem two stage design .....	41
2.4. Discussion .....	44
2.5. References .....	46
<b>Chapter 3. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review .....</b>	<b>64</b>
3.1. Abstract.....	<b>Error! Bookmark not defined.</b>
3.2. Introduction .....	64
3.3. Methods and Findings.....	65
3.3.1. Single Arm Designs .....	105
3.3.2. Enrichment Designs.....	106
3.3.3. Randomize-All Designs.....	114
3.3.3.1. Marker Stratified Designs .....	114
3.3.3.2. Hybrid Designs.....	136
3.3.4. Biomarker-Strategy Designs.....	137
3.3.4.1. Biomarker-Strategy Design with Biomarker Assessment in the Control Arm.....	138
3.3.4.2. Biomarker-Strategy Design without Biomarker Assessment in the Control Arm.....	142

3.3.4.3. Biomarker-Strategy Design with Treatment Randomization in the Control Arm.....	143
3.3.4.4. Reverse Marker-Based Strategy Design.....	147
3.3.5. Other Designs .....	150
3.3.5.1. A Randomized Phase II Trial Design with Biomarker Proposed by Freidlin et al., 2012 .....	150
3.4. Discussion .....	153
3.5. References .....	155
<b>Chapter 4. Online tool to help develop personalized medicine (BiGTeD) .....</b>	<b>171</b>
4.1. Introduction .....	171
4.2. Key features .....	172
4.3. User interface.....	173
4.3.1. Homepage .....	173
4.3.2. Design-specific webpages: Adaptive Designs .....	174
4.3.3. Design-specific webpages: Non-Adaptive Designs .....	178
4.4. Discussion .....	183
<b>Chapter 5. Fixed and adaptive Parallel Subgroup-Specific design for survival outcomes: power and sample size.....</b>	<b>186</b>
5.1. Introduction .....	186
5.2. Methods and Findings.....	187
5.2.1. Parallel Subgroup-Specific design.....	187
5.2.1.1. Sample size calculation for time-to-event-outcomes .....	189
5.2.1.2. Simulation Study 1 .....	192
5.2.1.3. Results from simulation study 1 .....	194
5.2.2. An adaptive version of the Parallel Subgroup-Specific design .....	199
5.2.2.1. Simulation Study 2.....	206



5.2.2.2. Results from simulation study 2 .....	208
5.3. Discussion .....	216
5.4. References .....	218
<b>Chapter 6. Deciding on the best biomarker-guided trial design: a case study ....</b>	<b>220</b>
6.1. Background to the proposed trial (STRONG trial) .....	220
6.1.1. Previously proposed randomized controlled trial of a stratified approach to Naltrexone treatment .....	226
6.1.2. Sample size calculation of the previously proposed design .....	229
6.2. Reasons for inefficiency of the previously proposed design .....	233
6.3. Reconsideration of the most appropriate design.....	234
6.4. Sample size calculations for the STRONG trial assuming different study designs .....	238
6.4.1. Using the Marker Stratified design in the STRONG trial.....	238
6.4.1.1. Using a variation of the Marker Stratified design – the Sequential Subgroup-Specific design .....	245
6.4.1.2. Using a variation of the Marker Stratified design: the Parallel Subgroup-Specific design .....	248
6.4.2. Using the Reverse Marker-Based strategy design in the STRONG trial .	253
6.5. Discussion .....	257
6.5.1. Non-adaptive design with time-to-event outcome .....	262
6.6. References .....	264
<b>Chapter 7. Case study - An adaptive approach.....</b>	<b>270</b>
7.1. Choosing the type of Sample size re-estimation method .....	270
7.2. Simulation study 1: With the option of early stopping of the trial for efficacy .....	272
7.2.1. Calculation of required sample size .....	272

7.2.2. Target number of patients at the interim stage.....	273
7.2.3. Variance of outcome .....	273
7.2.4. Test statistic of the first stage.....	274
7.2.5. Stopping boundaries.....	275
7.2.6. Sample size adjustment.....	277
7.2.7. Test statistic of the second stage .....	278
7.2.8. Final test statistic .....	279
7.2.9. Simulation parameters .....	279
7.3. Simulation study 2: With the option of early stopping of the trial for efficacy and futility.....	285
7.4. Simulation results .....	286
7.4.1. Binary Outcome.....	286
7.4.1.1. Control of type I error rate.....	286
7.4.1.2. Power of the study .....	288
7.4.1.3. Sample size of the study.....	290
7.4.1.4. Efficacy stopping probability .....	293
7.4.1.5. Futility stopping probability .....	296
7.4.2. Time-to-event Outcome .....	298
7.4.2.1. Control of type I error rate.....	298
7.4.2.2. Power of the study .....	303
7.4.2.3. Sample size of the study.....	309
7.4.2.4. Efficacy stopping probability .....	320
7.4.2.5. Futility stopping probability .....	328
7.5. Discussion .....	334
7.6. References .....	335
<b>Chapter 8. Challenges in Practice.....</b>	<b>337</b>

8.1. Examples of clinical trials .....	337
8.2. Challenges .....	340
8.2.1. Funding issues .....	340
8.2.2. Ethical and Regulatory Issues .....	342
8.2.3. Recruitment.....	345
8.2.4. Monitoring samples and labs .....	347
8.2.5. Biomarker assessment .....	348
8.2.6. Data sharing issues .....	349
8.2.7. Resources.....	350
8.3. Recommendations to overcome practical challenges .....	351
8.4. Discussion .....	353
8.5. References .....	354
<b>Chapter 9. Conclusions and Future Research .....</b>	<b>355</b>
9.1. Future directions .....	359
9.2. References .....	360
<b>Appendix A .....</b>	<b>362</b>
A.1. Variations of biomarker-guided adaptive trial designs .....	362
A.2. Literature review search strategies for both biomarker-guided clinical trial designs and for traditional trial designs .....	378
<b>Appendix B.....</b>	<b>381</b>
<b>Appendix C .....</b>	<b>386</b>
C.1. Supporting tables and figures .....	386
C.2. R codes .....	404
C.2.1. Parallel Subgroup-Specific design .....	404
C.2.2. An adaptive version of the Parallel Subgroup-Specific design .....	412
<b>Appendix D .....</b>	<b>418</b>

D.1. Results for Simulation study 1 .....	419
D.2. Results for Simulation study 2 .....	433
D.3. R codes.....	447
D.3.1. Simulation study 1 .....	447
D.3.2. Simulation study 2 .....	474
D.3.3. Type I error probabilities .....	501

# Chapter 1. Introduction

---

## 1.1. Personalized medicine

---

The idea of personalized medicine first appears in Hippocrates' time. Hippocrates of Cos, an ancient Greek physician who is considered the father of medicine said, "It's far more important to know what person the disease has than what disease the person has" [1]. Hence, the concept that a treatment will benefit patients who share similar characteristics in the same way and that patients can be treated with different drugs depending on their individual characteristics is not new and was well understood hundreds of years ago.

Several definitions have been provided for the term "personalized medicine", however, they all correspond to the idea of "one treatment fits a particular subgroup" instead of "one size fits all", meaning that patients are divided into several subpopulations according to their personal characteristics with each subpopulation assigned a treatment believed to benefit them the most. The ultimate aim of personalized medicine is to maximize benefit and minimize risk by tailoring an individual's treatment according to their personal characteristics. These characteristics can be demographic such as age or gender, or biological such as genetic or other biomarkers. According to the National Cancer Institute, "a biomarker is a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease, which may be used to see how well the body responds to a treatment for a disease or condition" [3].

Demonstrating the efficacy of using a personalized approach to treatment requires clinical trials. This thesis is focused on trial designs for identification of biomarker-defined subgroups which will derive maximal clinical benefit from a given treatment. Such designs have drawn considerable attention in the field of medicine as they promise an improved benefit-risk ratio for patients and enhanced opportunities for drug development.

### 1.1.1. Biomarker-guided clinical trial designs

---

In much the same way as the effectiveness of a treatment must be proven before it is prescribed in clinical practice, the effectiveness of using biomarkers to help tailor treatment must also be demonstrated before the biomarkers are used to guide treatment in practice. Trials that test the effectiveness of a biomarker-guided approach to treatment are collectively known as ‘biomarker-guided clinical trials’. The concept of personalized medicine has become increasingly popular over recent years, and consequently many different biomarker-guided trial designs have been proposed in the literature. The varying designs reflect the large variation in types of biomarkers and their different properties, with different designs being appropriate in different scenarios. For instance, some designs are aimed at the identification of a particular biomarker-defined subgroup which will benefit from a specific treatment whilst other designs might be aimed at assessing the impact of different biomarkers in a single type of disease (e.g. cancer) or testing the effect of a treatment(s) on a single biomarker in a variety of a disease types.

### 1.2. Novel insight into biomarker-guided clinical trial designs

---

The scope of this PhD thesis is to provide novel insights into biomarker-guided clinical trial designs. Whilst a number of these designs have been proposed in the past decade, there is lack of knowledge and understanding of current designs by those working in the field of personalized medicine including clinicians, policymakers and researchers. Further, there is very little guidance on which biomarker-guided design is appropriate in a particular setting. Consequently, it is difficult for investigators embarking on biomarker-guided clinical trials to decide on the most appropriate design for their particular setting, and navigating the literature to determine this can be difficult.

With this in mind, a thorough and comprehensive review of the biomarker-guided trial designs proposed to date is undertaken in order to give an in-depth overview of the designs for researchers working in personalised medicine, focussing

particularly on providing them with clarity in definition, methodology and terminology. For ease of reading, the review is split into two chapters, one focussing on adaptive biomarker-guided trial designs (Chapter 2) and the other focussing on non-adaptive designs (Chapter 3). Both chapters include key information related to utility, methodology, benefits and limitations of each design and examples of real studies.

### 1.2.1. Creating a user-friendly, interactive web resource

---

To enhance the guidance on biomarker-guided clinical trials, we developed clear graphical representations of each trial design, with standardized formatting to allow easy comparison across various designs. To ensure that the guidance and graphical representations are easily accessible to stakeholders, an interactive web resource to host this information was also developed. Chapter 4 presents this online tool for designing biomarker-guided clinical trials (BiGTeD, [www.BiGTeD.org](http://www.BiGTeD.org)). This work is informed by the methodological review presented in Chapter 2 and 3. BiGTeD is a user-friendly online tool which provides an easily accessible resource to inform on the most optimal design when embarking on a biomarker-guided trial including easy to navigate graphical displays of the various trial designs.

Our review and resulting guidance serves to improve the understanding of biomarker-guided trial designs and provides a valuable and much-needed resource for those wishing to implement such trials. This, in turn, will expedite the development of personalized approaches to treatment for the improvement in healthcare.

### 1.2.2. Choosing the best trial design

---

One size does not fit all, and investigators should take into consideration the implications that each design might have before selecting one. Chapter 5 focuses on a popular non-adaptive trial design, Parallel Subgroup-Specific design, for which a fixed sample size is required in advance of the study. We also explore an adaptive

approach of the same design aiming to assess the efficiency of the study related to the cost and time of the trial in general. Several statistical and simulation techniques are developed for the calculation of the required number of patients to ensure sufficient power for these types of designs.

Additionally, Chapter 6 reconsiders the most appropriate design for a clinical trial previously proposed which was not feasible due to several statistical considerations including a very large sample size that was required. Various biomarker-guided non-adaptive trial designs are explored and compared to choose which of them is more suitable for the purpose of the above trial. To assist in making this decision, we develop a series of strategies that could be used in future to decide between different trial designs. Sample size calculations are presented for both binary and time-to-event outcomes.

In Chapter 7, we explore an adaptive version of the chosen design through simulation studies in order to address a degree of uncertainty that surrounds the assumed effect size.

### 1.2.3. Guidance on practical challenges

---

In Chapter 8, we address the key practical challenges when undertaking biomarker-guided clinical trials, with particular focus on their management and monitoring. These challenges have been raised and discussed during the ‘Biomarker-guided trials: challenges in practice’ workshop organized by the author of this thesis and the MRC Hubs for Trials Methodology Research Network’s Stratified Medicine Working Group (SMWG) at the University of Liverpool in London Campus, 15th March 2017. A list of recommendations are drawn to overcome several key challenges.

## 1.3. References

---

1. U.S. Food and Drug Administration. Paving the way for personalized medicine 2013 [cited 2015 10 Oct]. Available online:



<http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalizedMedicine/UCM372421.pdf>.

2. Institute NHGR. Personalized Medicine [accessed on 23 August 2017]. Available online: <https://www.genome.gov/glossary/index.cfm?id=150>.

3. Institute NC. Biomarker [accessed on 23 August 2017]. Available online: <https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45618>.

## Chapter 2. Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review

---

### 2.1. Introduction

---

The rapidly developing field of ‘personalized medicine’ [1], also known as ‘individualized medicine’, ‘stratified medicine’, or ‘precision medicine’ is allowing scientists to treat patients by providing them with a specific regimen according to their demographic or individualized genomic or biological characteristics, known as biomarkers [2]. The terms Personalized Medicine and Individualized Medicine often create confusion in literature as, in reality, the objective of this approach is to identify demographic- or biomarker-defined subgroups. Thus, as it still remains a population and not an individualized approach, the terms Stratified or Precision medicine are often considered to be more accurate.

The Biomarkers Definitions Working Group defined a biomarker to be “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [1, 3-6]. Biomarkers related to clinical outcome which are measured before treatment can be classified as either prognostic or predictive biomarkers. Prognostic biomarkers provide information regarding the likely progression of a disease without taking into account any specific treatment, whilst predictive biomarkers provide information about the patient’s outcome given a certain treatment, i.e. their likely response to the treatment [3, 6-33].

Prior to utilizing a patient’s biomarker information in clinical practice, it is necessary that they have been robustly tested in terms of analytical validity, clinical validity and clinical utility. Specifically, the first term refers to the ability of a genetic test to detect and measure the biomarker of interest in a repeatable and reproducible way, hence it can answer the question of whether or not we should trust the results of a specific biomarker test. Once a valid test has been developed, the degree to which

the biomarker can identify the patients with and without the target disease from the whole population is referred to as clinical validity. A biomarker will be used in clinical practice only if it improves patient's health; hence, clinical utility corresponds to the assessment of the effectiveness of the biomarker and showing the positive effects compared to the risks of doing harm [8, 12, 18, 24].

A number of Phase II and Phase III trial designs have been proposed for testing the clinical utility of predictive biomarkers and they can be broadly classified into adaptive and non-adaptive trial designs. As we enter the new era of personalized medicine, there is substantial need for novel trial designs which will (i) demonstrate cost benefits and minimize the required time to obtain conclusive results despite an increase in the number of subjects needed for the trial; (ii) avoid erroneous conclusions and (iii) be more ethical by giving patients more effective treatments. Whereas non-adaptive trial designs often result in large and costly Phase III trials of long duration, adaptive designs are becoming increasingly attractive in the context of biomarker-directed therapies as they allow for additional flexibility during the course of the trial.

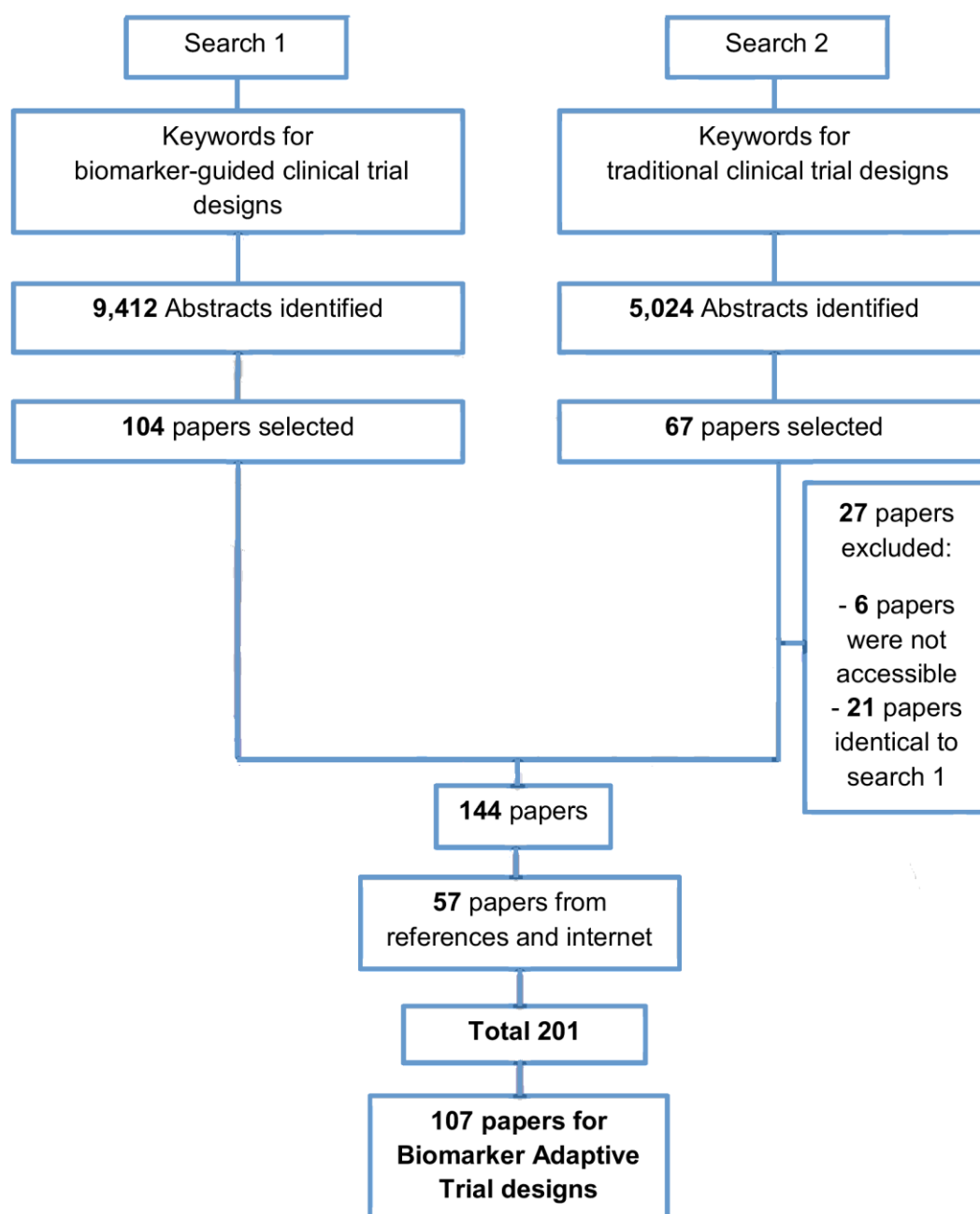
In the previous chapter (Chapter 1), we introduced the general concept of Personalized medicine and Biomarker-guided clinical trial designs. In the current chapter we report current biomarker-guided trial designs based on a comprehensive review in order to provide investigators embarking on such trials with useful guidance which to our best knowledge has not been performed previously to such an extent. This chapter is based on our published paper Antoniou et al. (2016) (see List of Publications) and is focused on adaptive trial designs. Key features are the graphical representation of each individual trial design with standardized formatting to allow easy comparison between them, categorization and presentation of key information including advantages and limitations.

## 2.2. Methods and Findings

---

We have undertaken a search of the MEDLINE (Ovid) database, restricted to published papers in the English language within the previous ten years aiming to identify articles which describe and discuss both non biomarker-guided trial designs, which we will refer to as ‘traditional’ trial designs, and biomarker-guided trial designs. Traditional trial designs are included in our literature review search strategy in order to ensure that we do not miss any potential reference to biomarker-guided designs, as the finding of the appropriate keywords in Medline database for biomarker-guided designs was challenging. Two separate strategies as illustrated in Figure 2.1 were used to identify relevant articles, and the keywords utilized in the search are presented in Appendix A.2. I screened the available titles and abstracts, and my supervisors, Dr Andrea Jorgensen and Dr Ruwanthi Kolamunnage-Dona were consulted when it was questionable whether or not a paper should be included. Our initial search resulted in 9,412 and 5,024 relevant titles for biomarker-guided clinical trial designs and traditional designs respectively. From the 9,412 papers, 104 articles were included based on their title and abstract. From the 5,024 papers, 40 articles were included based on their title and abstract and after removing inaccessible articles or those already identified in the search for biomarker-guided trial designs. A further 57 papers were identified from searching both the reference list of included articles and the internet (the internet searches were performed using the same keywords as those for the Ovid strategy). Of the 201 included papers, biomarker-guided adaptive trial designs were referred to in 107 papers. An extraction form was designed to collect all necessary information, and the summary of the extracted data was reviewed by my supervisors. If there were any ambiguities or confusion as to the extracted data, the second and third authors were consulted. For each included paper, the following details were extracted: definition of the trial design(s) referred to in the paper, how patients were screened and/or randomized based on their biomarker status, treatment groups randomized to, other key information relating to the trial design and methodology, advantages and limitations, and examples of actual trials which have adopted designs if mentioned together with

the proposed methodology and clinical field where the design had been applied. There is no evidence of some of the proposed trial designs being used in practice in the literature which was used for our review; however, they may well currently be in use in ongoing trials. The review of all trial designs which have been implemented in practice is beyond the scope of this chapter but if a published description of the methodology exists, this would have been captured by our search.



**Figure 2.1.** Flow diagram of the review process. Keywords are listed in Appendix A.2.

### 2.2.1. Adaptive designs

---

Before discussing the specific biomarker-guided adaptive trial designs, we consider key aspects of adaptive trial designs in more generality.

#### 2.2.1.1. *Definitions and terminology*

---

To date, several authors have given different definitions about adaptive designs in general [34-36]. Chow et al. (2005) [34] described the adaptive design as a strategy that allows adaptations in trial procedures and/or statistical procedures after initiation of the trial without undermining the validity and integrity of the trial. In 2006, the Pharmaceutical Research Manufacturer Association (PhRMA) Working Group on Adaptive Design [35] defined an adaptive design as a clinical trial design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial.

In 2010, US Food and Drug Administration defined an adaptive design as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of (usually interim) data from subjects in the study [36]. In the context of biomarker-guided therapies, Chen et al. (2014) [12] defines the biomarker adaptive trial design as “designs which identify most suitable target subpopulations with respect to a particular treatment, based on either clinical observations or known biomarkers, and evaluate the effectiveness of the treatment on that subpopulation in a statistically valid manner”.

Some researchers refer to these approaches as flexible designs [33, 37-40], terminology which can cause confusion since some trial designs which allow adaptivity are by no means flexible, for example those with pre-specified rules in terms of how to proceed based on interim data analyses [41]. Thus, the term ‘flexible designs’ can include designs with both planned and unplanned properties [42].

#### 2.2.1.2. *Adaptations to the design*

---

Adaptations based on interim analysis, which are made during the course of an adaptive strategy include adding or dropping treatment arms, changes in the required sample size, changes in the allocated proportion of the study population in order to randomize more patients to treatment arms which are doing better, or refinement of the existing study population according to their predictive biomarkers [38-40, 43-49].

In Personalized Medicine some common adaptations during the implementation of adaptive designs refer to changes in randomization probabilities within the biomarker-defined subgroups or dropping a biomarker-defined subgroup [15, 50].

Generally, this type of biomarker-guided approach is appropriate when (i) the candidate biomarker is not known at the start of the study; (ii) there are multiple experimental treatments and pre-specified biomarker-defined subgroups; (iii) existence of well-established analytical validity; (iv) rapid turnaround time for biomarker assessment [12, 15, 51]. Finally, it is important to have informative endpoints that are observed early.

#### 2.2.1.3. *Analysis of adaptive designs*

---

Although both a Bayesian and Frequentist framework has been used for the analysis of adaptive designs [26, 52-54] the former has been described by many authors as a more suitable perspective due to its flexibility as it enables revision of knowledge according to prior information. It is common that Bayesian methods are used but frequentist operating characteristics are controlled. I-SPY2 and BATTLE studies are examples of actual adaptive trials designed with a Bayesian framework [48, 49, 55]. Nevertheless, the Bayesian perspective in adaptive designs can cause many problems in terms of computational demands, inference making and parameter estimations [10, 26, 55, 56].

### 2.2.2. Biomarker-guided adaptive trial designs

---

In our review, we have identified eight main biomarker-guided adaptive designs. Four of the eight designs also have variations. Each main design is presented graphically in Figures 2.2-2.9. The characteristics and methodology of the eight main designs are discussed below and summarized in Table 2.1, whilst information on their variations are summarized in Table A.1 in Appendix A.1.



**Table 2.1.** Characteristics of biomarker-guided adaptive trial designs in Phase II and Phase III

Types of Biomarker-guided adaptive trial designs	Phase	Adaptations	Pros	Cons
<b><u>Adaptive signature design</u></b> <b>(30 papers)</b> [2, 6, 8, 9, 12, 14-16, 18, 20, 21, 24, 27, 31, 32, 47, 49, 57, 58-68]  <b>Also called:</b> Two-stage Adaptive Signature design, Adaptive Two-stage design, Biomarker-Adaptive Signature design	III	Identification of biomarker-positive subpopulation	Identification of optimal group of patients which benefit the most from a specific treatment.  Identification and validation of candidate biomarker signature in a single trial.  Avoids inflation of type I error rate as it does not use the individuals on which the predictive signature was developed in order to test the treatment effect.  Rapid and efficient approval of the novel treatment.  No modifications in randomization weights or in eligibility criteria are made, consequently, it	Larger sample size may be required, especially when there is small difference between biomarker-negative and biomarker-positive patients.  Can limit its power when testing the treatment effectiveness in the biomarker-positive subgroup as half of patients are used for signature development and half for validation of the biomarker.  Treatment comparisons can only be performed when the study is completed.

			avoids any statistical adjustment needed to ensure that there is no introduction of bias.	
<b><u>Outcome-based adaptive randomization design</u></b> (24 papers) [14, 26, 29, 32, 37, 40, 47, 49, 52, 56, 59, 62, 63, 65, 69, 70-78]  <b>Also called:</b> Adaptive randomization Bayesian Adaptive, Bayesian Adaptive randomization, Combined dynamic multi-arm, Outcome-Adaptive randomization, Outcome-based Bayesian Adaptive Randomization	II	Change in randomization ratio	Smart, novel, and ethical approach  Permits updating patient's outcome (it uses the accumulated information in order to assign patients to different treatment arms; the arm which seems to benefit the study population the most, is composed of the higher proportion of randomized individuals).  Can result in high probability of success of the trial as there is increase in the number of patients who receive effective treatments.  In the Bayesian perspective, Type I and II errors can be controlled by carefully calibrating the design parameters.  Can boost patients' ethics as patients are assigned to the best available therapy.	Complexity in terms of building-up the trial design, conduct and analysis of the trial.  Can make incorrect decisions in case of incorrect biomarker selection as the design is based on the accumulated data about how well the biomarker performs.  Requirement of relatively short biomarker and endpoint assessment (quick testing of the biomarker is required in order to avoid incorrect decision regarding the assignment of patients and rapid assessment of outcome to randomize adaptively according to the updated outcome.).  Likely to introduce bias due to time trends in the prognostic mix of individuals

				enrolled to the study. The evidence accumulated during the trial can influence clinicians' decisions regarding which patient groups to consider recruiting.
<b><u>Adaptive threshold sample-enrichment design</u></b> (5 papers) [18, 20, 21, 63, 79]  <b>Also called:</b> Threshold sample-enrichment approach, Two-stage Sample Enrichment, Two-stage sample-enrichment design strategy	III	Change in the inclusion criteria of the study population after the initial stage of the study in order to broaden the targeted patient population.	<p>More cost-effective as it avoids further recruitment of patients when there is no difference in treatment outcome among the biomarker-defined subgroups.</p> <p>Researchers can use the data which was accumulated during the first stage of the study to proceed with further investigation of any other potential assumption made at the start of the trial.</p>	Cannot work if there is no information about a subset of patients for whom the novel treatment is more effective than others before the beginning of the trial.
<b><u>Adaptive patient enrichment design</u></b> (23 papers) [3, 6, 7, 14, 18, 20, 21, 29, 38, 42, 43, 63, 70, 74, 78, 80-87]	III	Information obtained from interim stage is used to decide whether the study should be narrowed	Can detect a particular biomarker-defined subgroup most likely to respond to the novel treatment, thus the efficiency of study design can be increased.	<p>Can be quite complex.</p> <p>Can result in biased treatment effect estimates.</p> <p>Criticised as a design without satisfactory operating characteristics in real practice</p>

<b>Also called:</b> Adaptive accrual, Adaptive accrual based on interim analysis design, Adaptive Enrichment, Adaptive Modification of Target Population. Adaptive Population Enrichment, Two-stage Adaptive Design, Two-stage adaptive accrual	to a particular subpopulation.	Can gain that more power than a fixed study design under the scenario that the genomic biomarker is predictive of treatment effect (i.e. the value of effect size indicates that there is treatment effect in the biomarker-defined subgroup, e.g. the value of 0.4) than in the case where the genomic biomarker is prognostic (i.e. the scenario where we assume that the value of effect size is zero).	<p>with a lack of generalizability and information in subgroups which are excluded.</p> <p>May augment the duration of the trial depending on the prevalence of the biomarker for the biomarker-defined subgroup which continues to full accrual due to recruitment of many more biomarker-positive patients.</p> <p>Requirement of an appropriate futility boundary and rapid and reliable clinical endpoint.</p> <p>Conservativeness of futility boundaries as the futility boundary is set to be in the region in which the observed efficacy of the standard of care is greater than that for the experimental treatment.</p>
---	--------------------------------	--	---

				Assumes complete confidence in the biomarker.
				Early termination of the entire trial is not permitted.
<b><u>Adaptive parallel Simon two-stage design</u></b> (8 papers) [6, 76, 85, 88-92]  <b>Also called:</b> Biomarker-adaptive parallel two-stage, Adaptive parallel, Two-parallel Simon, Two-stage design	II	The design starts with two parallel studies and according to the results of the initial stage we enrol selected or unselected patients during the second stage.	May reduce the required sample size.  May augment the efficiency of the trial as it allows for early understanding that a particular experimental treatment is beneficial in a specific biomarker-defined subgroup.  Straightforward and simple strategy with reasonable operating characteristics.	Does not allow early termination of the trial for efficacy in biomarker-defined subgroups during the first stage of the trial.
<b><u>Multi-arm multi-stage designs</u></b> (16 papers) [18, 20, 40, 56, 63, 69, 89, 93-103]  <b>Also called:</b> Adaptive biomarker-driven design, Adaptive Analysis, Adaptive	II/III	Experimental arms can be dropped for futility from the study.	Promising treatments are tested concurrently using a smaller number of patients as some treatments arms can be dropped early for futility.	High setting-up time due to the complexity caused by logistic issues and collection of experimental drugs from different companies.  Operational challenges regarding the randomization and the modifications of

<p>Multi-stage designs, Multi-stage</p>	<p>Reduced costs and time as they assess multiple treatments simultaneously.</p> <p>Preferable to continue with the investigation of promising treatments as compared to the conduct of separate single-arm phase II clinical trials.</p> <p>The simultaneous assessment of multiple experimental treatments increases the chance of identifying a promising treatment.</p> <p>It is unlikely that the trial will stop for futility as multiple experimental treatments are tested and hence, it is not likely that all experimental arms will be ineffective and dropped.</p> <p>Can ease the regulatory and administrative burden as compared to building-up separate trials.</p>	<p>allocation ratios after the performance of an interim analysis.</p>
---	---	--

Unpromising experimental arms can be dropped in a quick and reliable way.				
<b><u>Stratified adaptive design</u></b> (1 paper) [89]	II	The number of patients and decision rules are based on the observed response rates during the first stage of the study.	Can avoid unethical studies in patients for whom the novel treatment is not effective as it allows for the identification of efficacy which is limited to a particular biomarker-defined subgroup.  The trial can continue to Phase III only with a subgroup which is proven to benefit from the experimental therapy and not with the entire population.  Less numbers of individuals for whom the novel treatment is not effective will be tailored to toxic treatments.  Permits the identification of the actual treatment benefit in at least one biomarker-defined subgroup.	No information found
No alternative names found for this trial design				

		<p>Avoids the termination of tailoring a novel treatment due to treatment effect dilution in the entire population.</p> <p>Permits early stopping of efficacy or inefficacy.</p>		
<p><b><u>Tandem two stage design</u></b> (5 papers) [21, 63, 76, 90, 104]</p> <p><b>Also called:</b> Tandem two-step phase II trial, Tandem-two step trial (phase II), Tandem two-step phase 2 trial design, Tandem two-step</p>	II	<p>Assessment of treatment effectiveness in the entire population at the first stage of the study to make a decision about enriching the targeted patient population.</p>	<p>Although the two stages could be run separately, i.e. one for the biomarker-positive subgroup and the other for the unselected patients, the performance of the study in this way can increase the duration and costs of the trial. Consequently, it will be better to run the study in just one trial so as to have a more seamless study.</p> <p>Allow estimating response rates not only in the unselected biomarker-defined patients (entire population) but also in the biomarker-positive subset.</p> <p>Identify whether the experimental treatment is beneficial in the entire population, and if it is not, then can test whether the candidate</p>	No information found



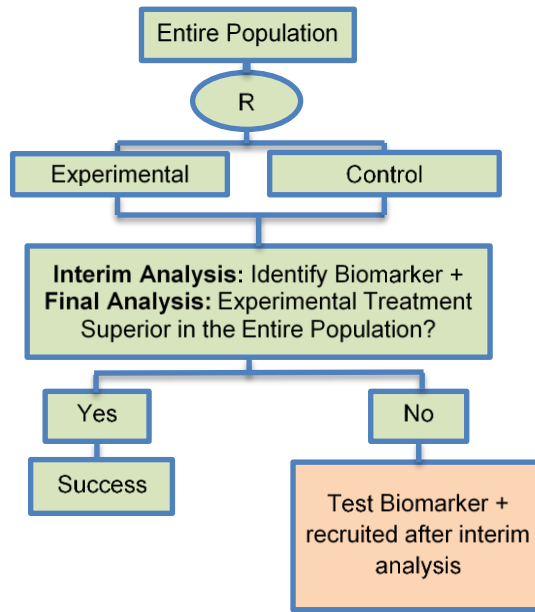
predictor can enrich the responding population.

Allow for simultaneous testing of multiple different biomarkers for the same treatment in a single parallel multi-arm trial.

The adaptive signature design is described in 30 (28%) papers of our review. It is a two-stage Phase III non-Bayesian trial design proposed by Freidlin and Simon (2005) [105] for settings where an assay or signature that identifies sensitive patients (i.e. biomarker-positive patients) is not known at the outset. This trial design permits the development and evaluation of a biomarker based on high dimensional data. It uses a training set to identify predictive biomarkers and evaluates them in a validation set. Generally, this approach is useful when there is no available biomarker at the start of the trial or there is a great number of candidate biomarkers which could be combined to identify a biomarker-defined subgroup, and the attention is given first to the entire study population. Five variations of the adaptive signature design have also been identified, with differences occurring mainly in terms of the analysis. These variations are the following: i) Adaptive threshold design, ii) Molecular signature design, iii) Cross-validated Adaptive Signature design, iv) Generalized adaptive signature design and v) Adaptive signature design with subgroup plots. Information about each variant can be found in Appendix A.1, section “Variations of Adaptive signature design”.

**Design:** Figure 2.2 graphically represents the trial design. The design begins with a comparison between the experimental treatment and the standard treatment in the entire study population at a pre-specified level of significance. In case that the overall result is positive, it is considered that the treatment is beneficial, and the trial is closed. If the comparison in the overall population is not promising, then the entire population is divided in order to develop and validate a biomarker, using a split sample strategy. More precisely, a portion of patients is used to detect a biomarker signature that best distinguishes subjects for which the novel treatment is better than the standard treatment. Hence, this approach (i) identifies patients who are more susceptible to a specific treatment during the initial stage of the study (at the interim analysis); (ii) it assesses the global treatment effect of the entire randomized study population through a powered test, and (iii) finally, it assesses the treatment effect for the biomarker-positive subgroup identified during the initial stages of the study

but only with patients randomized in the remainder of the trial, the so-called ‘validation test’.



**Figure 2.2.** Adaptive signature design. “R” refers to randomization of patients.

**Methodology:** The analysis is undertaken as follows: If the overall treatment effect is not significant at a reduced level  $\alpha_1$  ( $< 0.05$ ), the full set of  $P$  patients in the clinical trial is partitioned into a training set  $Tr$  (recruited before interim analysis) and a validation set  $V$  (recruited after interim analysis). A pre-specified algorithmic analysis plan is applied to the training set to generate a classifier  $Cl(x;Tr)$  where  $x$  is a biomarker vector. This classifier is a function that identifies a biomarker-positive subgroup of patients who appear to benefit from the experimental treatment  $E$ .  $Cl(x;Tr)=1$  means that a patient with  $x$  is predicted to benefit from  $E$  whereas  $Cl(x;Tr)=0$  indicates that a patient is not predicted to benefit from  $E$  [57]. The experimental treatment  $E$  is compared with the standard of care (or control) treatment in the biomarker-positive subgroup of the validation set using a significance level of  $\alpha_2 = \alpha - \alpha_1$  in order to ensure that the overall likelihood to obtain a false-positive conclusion is no greater than  $\alpha$  ( $= 0.05$ ).

Freidlin and Simon (2005) [105] recommended that a level of  $\alpha_1 = 0.04$  (two-sided) is allocated to the entire population hypothesis and  $\alpha_2 = 0.01$  (two-sided) is allocated to the biomarker-positive subgroup hypothesis. The multiplicity problem is a concern with this approach as the statistical test is conducted twice; this is why the alpha is split so the total is 0.05. The power of this strategy can be increased using K-fold cross-validation as Freidlin and Simon (2005) [105] demonstrated (see the Cross-validated adaptive signature design (CVASD) in Appendix A.1 for further information).

**Statistical/practical considerations:** Although the adaptive signature design allows for approval of the novel treatment in a quick and efficient way as it combines two trials into one, the main statistical challenges to be taken into account include the potential increase in the number of patients and the limited power to assess the treatment effect in the biomarker-defined subgroup. However, this approach avoids introduction of bias since the adaptations do not involve modifications in allocation ratio and eligibility criteria. Further, it prevents the inflation in type I error rate as the design does not use the study population which was employed to develop the predictive signature for the assessment of the treatment effect.

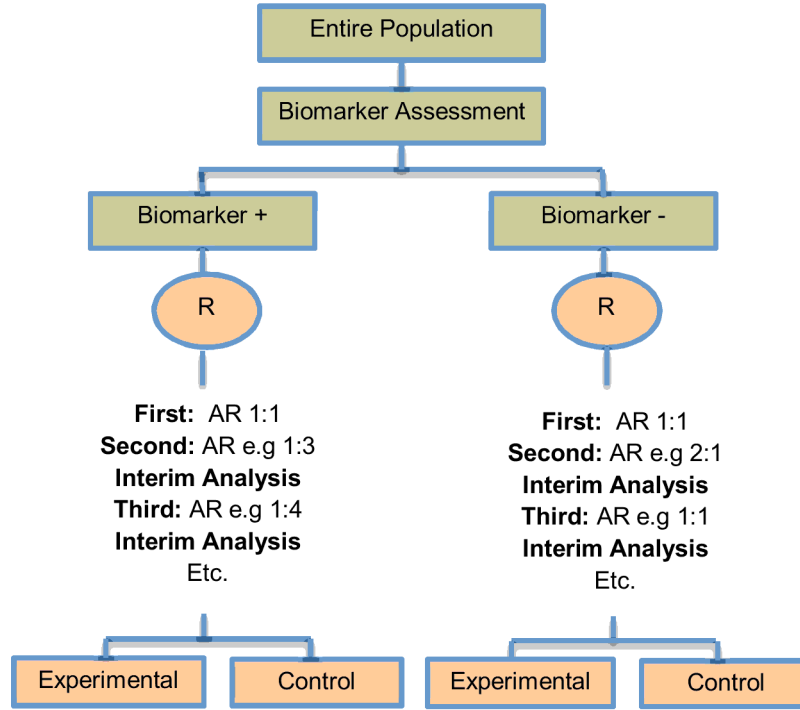
#### *2.2.2.2. Outcome-based adaptive randomization design*

---

The outcome-based adaptive randomization approach is referred to in 24 (22.4%) papers. In the context of personalized medicine, this design is used when the biomarkers are only putative or not known at the beginning of a Phase II trial and is also useful when there are multiple targeted treatments and biomarkers to be considered. It aims to test simultaneously both biomarkers and treatments while providing more patients with effective therapies according to their biomarker profiles. Outcome-adaptive randomization is sometimes included under the umbrella of “Bayesian clinical trials” but as criticized by Korn and Freidlin (2011) [71], there is nothing inherently Bayesian about it. There is a single variant of the Outcome-based adaptive randomization design with differences occurring in its analysis methodology. This variant is referred to as Bayesian covariate adjusted

response-adaptive randomization and information about this approach can be found in Appendix A.1, section “Variation of Outcome-based adaptive randomization design”. Two examples of actual trials which use the outcome-based adaptive randomization approach are the following: i) BATTLE trial mentioned in [14, 29, 52, 59, 62, 70, 72-74, 76, 77], ii) ISPY2 mentioned in [29, 32, 49, 62, 72, 75, 76].

**Design:** An illustration of this approach with one biomarker is shown in Figure 2.3. The trial begins with the assessment of patients’ biomarker status. The design permits the modification of patients’ allocation to different treatment arms so that the arm(s) which seem(s) to benefit the study population the most, is composed of the higher proportion of randomized patients. Consequently, we have randomization probabilities which do not stay fixed over time (e.g. change from adaptive randomization (AR) ratio 1:1 to AR 2:1, see Figure 2.3). The random assignment of patients to treatment arms, according to their biomarker status, depends on the use of accumulated patients’ data about how well the biomarker performs as at each interim analysis stage. When these accruing outcome data indicate that an experimental treatment is more effective as compared to the standard of care (or other treatment(s) or control), it is possible that a higher number of patients will be assigned to this particular experimental arm.



**Figure 2.3.** Outcome-based adaptive randomization design. “R” refers to randomization of patients.

**Methodology:** Zhou et al. (2008) [77] proposed an analysis plan in a Bayesian hierarchical framework using the Bayesian probit model to characterize the disease control rate for each treatment by biomarker subgroup. Therefore, the estimates for the treatment and the biomarker effect are provided by using the adaptive randomization design with the incorporation of a hierarchical Bayes model (it is a probit model included in the category of generalized linear models which uses the probit link function to model categorical or ordinal data). More precisely, the process starts with the biomarker profile assessment of all eligible patients and then according to the profile of each individual, the study population will be assigned to the different biomarker groups (e.g. a patient with a particular biomarker will be assigned to a specific biomarker group). Due to the fact that at the beginning of the trial we do not know the true disease control rate (i.e. the proportion of patients who demonstrate response to a treatment) the trial begins with equal randomization so that each treatment by biomarker subgroup is composed of at least one individual with a known disease control status (whether the patient will experience progression given a certain treatment). Next, the trial continues with adaptive randomization of patients; this is achieved by using the Bayesian probit model to calculate the posterior

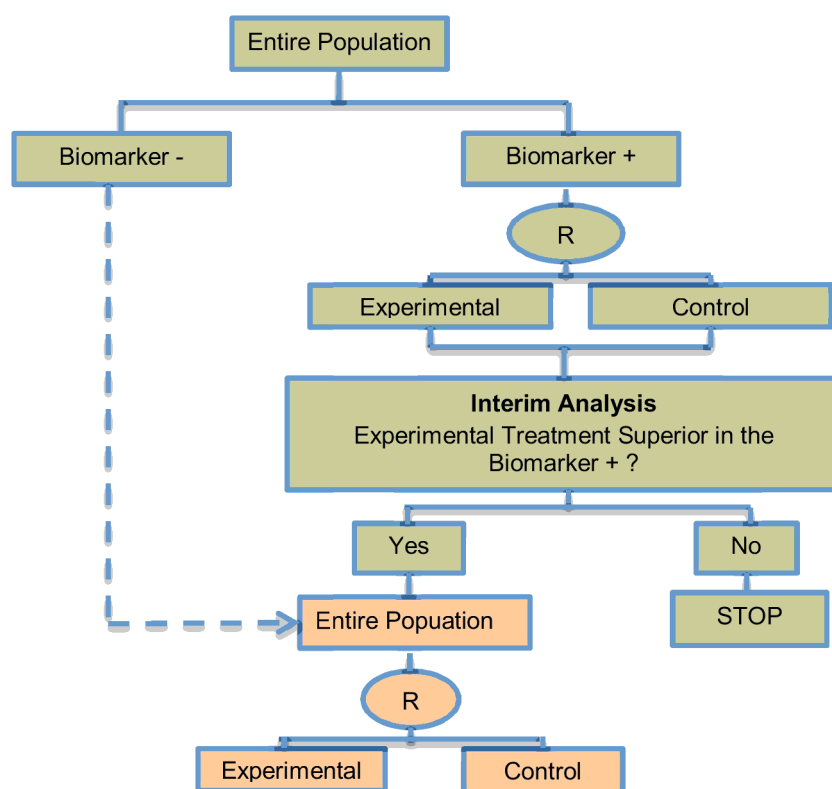
disease control rate. After the posterior rate is found, we define the randomization rate as the posterior mean of the disease control rate of each treatment in each biomarker-defined subgroup. The adaptive randomization process continuous until the last individual is enrolled and can stop early only in case that all treatments are dropped due to inefficacy. Whereas in many trial designs the baseline covariate (in this case the biomarker) is considered as prognostic, the design proposed by Zhou et al. (2008) [77] allows for modelling the treatment by biomarker interactions where the biomarker is in fact predictive. The incorporation of the above hierarchical Bayesian structure allows ‘borrowing strength’ or information-sharing across patients receiving the same treatment but with different biomarker profiles [77].

**Statistical/practical considerations:** Despite the fact that this design can be considered successful as an ethical approach where patients can be assigned to the most effective treatments according to their biomarker profiles, an issue that raises concern is the requirement of a relatively short assessment period of both biomarker and endpoint to avoid erroneously not only the assignment of patients but also the adjustment of the randomization rate. Also, potential introduction of bias due to time trends in the prognostic mix of the patients enrolled to the study should be taken into consideration.

#### *2.2.2.3. Adaptive threshold sample-enrichment design*

---

Adaptive threshold enrichment design was identified in 5 papers (4.7%) of our review. This approach is a two-stage design in a Phase III setting which was proposed by Liu et al. (2010) [79] to adaptively modify accrual in order to broaden the targeted patient population (see Figure 2.4).



**Figure 2.4.** Outcome-based adaptive randomization design. “R” refers to randomization of patients.

**Design:** The design is based on the former knowledge that a specific biomarker-defined subgroup (biomarker positive) is believed to benefit more from a novel treatment as compared to the remainder of the study population (biomarker negative). This knowledge can be gained, for example, from previous studies such as a large scale comparative trial (Phase III) when there was treatment effect heterogeneity among the study population. This design allows the trial to be terminated for futility in the biomarker positive subgroup. More precisely, the trial proceeds as follows: (i) accrue and randomize only biomarker positive patients; (ii) conduct an interim analysis in order to compare the experimental treatment with the standard of care within the biomarker positive subgroup; (iii) if the interim result is negative, then the accrual stops and the trial is closed without showing a treatment benefit; if the result is ‘promising’ for the specific biomarker-positive subgroup, then the study continues with this specific biomarker positive subgroup and accrual also begins for biomarker negative patients. Thus, the trial continues with patients randomized from the entire population. A ‘promising’ result in the biomarker



positive subgroup at the interim stage is claimed when the estimated treatment effect is above a particular pre-specified threshold.

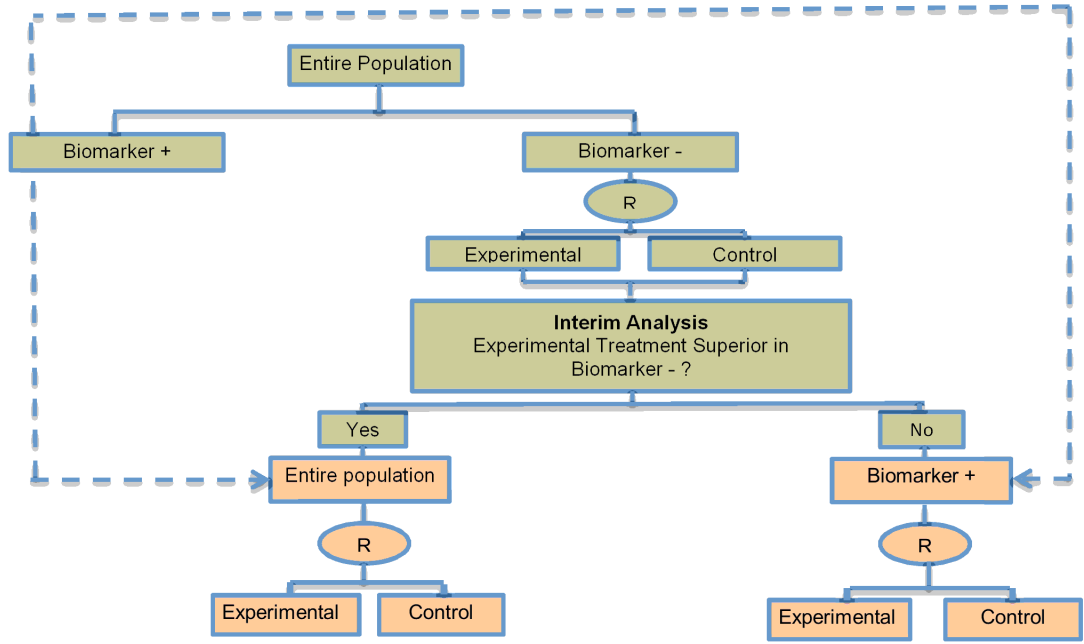
**Methodology:** The analysis is undertaken as follows: At the interim analysis stage, the treatment effect of a sample of patients ( $n_1$ ) from the biomarker-positive subset is estimated. If an improvement is seen in the experimental treatment arm which is greater than a pre-specified threshold value (i.e. the estimated treatment difference between the novel treatment arm and the control treatment arm for this subpopulation is greater than a threshold value  $c$  divided by the square root of the aforementioned sample size  $n_1$ ) the trial continues with accrual of patients from the entire biomarker-positive subgroup and additional patients are also accrued from the biomarker-negative subpopulation; otherwise the trial is stopped for futility. At the end of the trial, the treatment effect is estimated for all subpopulations. Researchers should choose the sample size  $n_1$  so that a persuasive result can be reached when the first stage of the trial is completed. In general, the threshold value  $c$  can be determined so that  $c/\sqrt{n_1}$  is a proportion of the smallest meaningful treatment improvement that researchers expect, e.g. it can be set to half of the smallest clinically important difference. Other methods also have been proposed [79].

Liu et al. (2010) [79] give a detailed description regarding the statistical formalization of the Type I error rate of this two-stage test and the power for assessing group-specific treatment effects. Also, Liu et al. (2010) [79] give detailed information on testing hypotheses based on the overall treatment effect indexed as a weighted average of the group-specific treatment effects, where the weight can be specified as the prevalence of that particular subgroup.

**Statistical/practical considerations:** The Adaptive threshold sample-enrichment design is not feasible if there is no prior knowledge regarding a subgroup of patients which is more susceptible to a particular treatment than others. In addition, this approach is considered more cost-effective as there will be no further recruitment from the study population when there is no evidence of treatment effectiveness.

The adaptive patient enrichment design was included in 23 papers (21.5%). This is a two-stage Phase III clinical trial design proposed by Wang et al. (2007) [80]. There is a single variant of the adaptive patient enrichment design with differences occurring in its methodology. This variant is referred to as Modified Bayesian version of the two-stage design of Wang et al. (2007) [80] and information about it can be found in Appendix A.1, section “Variation of Adaptive patient enrichment design”. An example of actual trial which incorporates the adaptive patient enrichment design is the NCT01001234 trial [42, 87].

**Design:** This approach is used for the comparison of an experimental treatment with the standard of care (control) which adaptively modifies accrual to two predefined biomarker-defined subgroups based on an interim analysis for futility. Figure 2.5 presents the adaptive patient enrichment design, and in general it flows as follows: (i) accrue both biomarker-positive and biomarker-negative patients; (ii) perform an interim analysis to evaluate the experimental treatment in the biomarker-negative subgroup; (iii) if the interim result in that subgroup is ‘not promising’, defined as the observed efficacy for the control group being greater than that for the experimental group and the difference being greater than a futility boundary, then accrual of biomarker-negative patients stops; but the strategy continues with accruing additional biomarker-positive patients in order to substitute the unaccrued biomarker-negative patients until the pre-specified total target sample size is achieved; (iv) contrarily, if the interim results are promising in the biomarker-negative patients, the accrual of both biomarker-negative and biomarker-positive patients continues until the total target sample size is achieved.



**Figure 2.5.** Adaptive patient enrichment design. “R” refers to randomization of patients.

**Methodology:** A pre-planned total sample size with futility stopping is considered for this two-stage adaptive design. The trial assesses the treatment effect both in the entire population and in the biomarker-positive population. Wang et al. (2007) [80] performed a simulation study testing a composite hypothesis; the hypothesis of the global treatment effect and a hypothesis of treatment effect in the biomarker-positive subgroup. A bivariate normal model which incorporates the correlation between the two test statistics for each hypothesis was used. Furthermore, two multiplicity adjustment methods which have a strong control of experimentwise false-positive rate ( $\alpha = 0.025$ ) were considered in order to test the composite hypotheses of no treatment effect; the first method was the equal split-alpha method which allocate  $\alpha_1 = \alpha_2 = 0.0125$  [106] and the second method was the Hochberg’s method [107] for multiple testing; a special case of partitioning  $\alpha$  which starts with the least significant p-value and investigate the other p-value in a sequential manner until it reaches the most significant one (unequal alpha split).

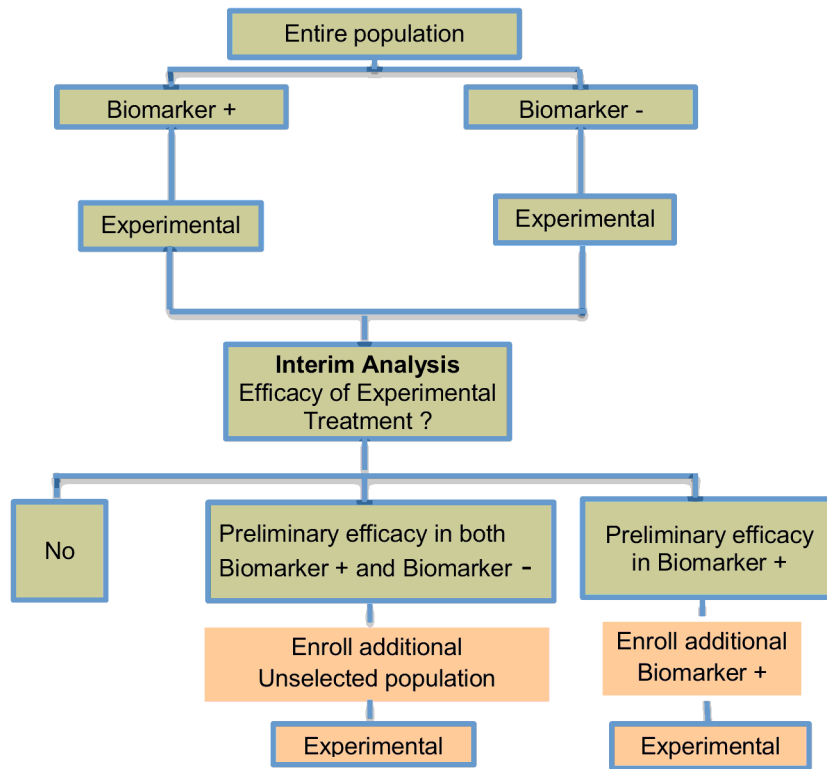
**Statistical/practical considerations:** Although a greater power is achieved as compared to a non-adaptive trial design (i.e. fixed study) in simulation settings, this strategy can yield an important increase in the duration of recruitment depending on the prevalence of the biomarker. Additionally, it does not allow for early termination

due to efficacy and can lead to biased treatment effect estimates when the results from interim analysis are used for exclusion of a biomarker-defined subgroup. In addition, this study design is appropriate when there is rapid outcome assessment relative to the accrual rate and assumes complete confidence in the biomarker at the outset. A further limitation is that the futility boundary is considered conservative and less than optimal. This conservative approach was chosen in order to expose the least number of patients possible to a non-beneficial treatment. However, alternative options could be considered.

#### *2.2.2.5. Adaptive parallel Simon two-stage design*

---

Jones and Holmgren (2007) [85] proposed a Phase II adaptive design (Figure 2.6) by extending the Simon two-stage design [88]. This strategy does not include a control arm yet, consequently it can be considered a single-arm approach exactly like the Simon two-stage approach. The biomarker-adaptive parallel Simon two-stage design was mentioned in 8 (7.5%) papers of our review. The design aims to test a novel treatment which possibly has a different treatment effect in the biomarker-positive versus the biomarker-negative subgroups. This approach requires a pre-defined biomarker with well-established prevalence and permits preliminary determination of efficacy that may be restricted to a particular subset of patients. An example of actual trial which uses this strategy is the NCT00958971 trial [76, 92, 108].



**Figure 2.6.** Adaptive parallel Simon two-stage design. “R” refers to randomization of patients.

**Design:** The design begins with two parallel phase II studies. During the first stage, two separate studies are performed in the biomarker-positive and biomarker-negative subgroups. Next, depending on the interim results of the first stage, the trial either stops or continues into a second stage with the enrollment from either the entire patient population (unselected patients) or from the biomarker-positive subpopulation only (selected patients). If a preliminary efficacy is observed during the first stage of the study for the experimental treatment in both the biomarker-positive and biomarker-negative subset, then additional patients from the general patient population will be enrolled in the second stage; if the interim result during the first stage of the trial shows that the efficacy is limited to the biomarker-positive subjects, then the recruitment of additional biomarker-positive patients only continues during the second stage.

**Methodology:** If there are sufficient results in both first and second stages, the novel treatment can further be explored. More precisely, the strategy is as follows as outlined by McShane et al. (2009) [90]: In the first stage of the design,  $N_1^-$  biomarker-

negative individuals and  $N_1^+$  biomarker-positive individuals are recruited. An interim analysis is undertaken with its results determining how the design proceeds as follows: If the number of responses to the novel treatment in the biomarker-negative group, in the first stage  $X_1^-$ , meets or exceeds a cutoff of  $k_1^-$ , then  $N^{un}$  additional unselected individuals are accrued during the second stage (including  $X_2^-$  biomarker-negative responders and  $X_1^+$  biomarker-positive responders). If  $X_1^-$  is less than  $k_1^-$  but the number of responses in the biomarker-positive group in the first stage,  $X_1^+$ , meets or exceeds a cutoff of  $k_1^+$ , then the design enrolls  $N_2^+$ , additional biomarker-positive individuals during the second stage (including  $X_2^+$  responders). If  $X_1^-$  is less than  $k_1^-$  and  $X_1^+$  is less than  $k_1^+$  then the trial stops. Consequently, when the second stage is terminated, a total of  $N^+$  and  $N^-$  biomarker-positive and biomarker-negative individuals, respectively, will have been enrolled, whilst a total of  $X_T^+$  (biomarker-positive group) and  $X_T^-$  (biomarker-negative group) responders will have been observed.

At the end of stage two, treatment benefit is determined as follows: In the case where unselected individuals continued to be accrued during the second stage, the total number of responders in the biomarker-negative subgroup,  $X_T^-$ , is compared to a cutoff,  $k^-$  whilst the total number of responders in the biomarker-positive subgroup,  $X_T^+$ , is compared to a cutoff,  $k^+$ . If  $X_T^- \geq k^-$ , then we conclude that the experimental treatment is beneficial in the unselected population; if  $X_T^+ \geq k^+$  and  $X_T^- < k^-$  then we conclude that the treatment is beneficial only in the biomarker-positive population; if  $X_T^+ < k^+$  and  $X_T^- < k^-$ , then we conclude no treatment benefit. In the case where only biomarker-positive patients continued to be accrued during the second stage,  $X_T^+$ , is compared to a cutoff,  $k_2^+$ . If  $X_T^+ \geq k_2^+$  then we conclude treatment is beneficial in the biomarker-positive subgroup; otherwise we conclude no treatment benefit. The trial stage- and subgroup-specific sample sizes  $N_1^-, N_1^+, N^{un}, N_2^+$  and cutoffs  $k_1^-, k_1^+, k^-, k^+, k_2^+$  are determined so that they control the probability of correct conclusions in the biomarker-positive and unselected patient groups.

Jones and Holmgren (2007) [85] have used the values 34, 34, 32 and 36 for  $N_1^-$ ,  $N_1^+$ ,  $N^{un}$ , and  $N_2^+$  respectively and the values 2, 1, 4, 4 and 5 for  $k_1^-$ ,  $k_1^+$ ,  $k^-$ ,  $k^+$ , and  $k_2^+$  respectively. As stated by Jones and Holmgren (2007) [85] values for the cutoffs  $k_1^-$  and  $k_1^+$  (equal to 2 and 1 respectively) are obtained from the first stage of the optimal Simon two-stage design. Additionally, in the case where there is preliminary efficacy of the experimental treatment in the unselected population during the first stage of the trial, the study enters the second stage where the values of  $k^-$  and  $k^+$  for decision making need to be defined. Assuming the total number of biomarker-positive subjects ( $N^+$ ) enrolled by the end of the second stage is fixed at its expected value given a known prevalence, the aforementioned values ( $k^-$  and  $k^+$ ) can be acquired as the minimum values needed for exclusion of the null value from the  $(1 - \alpha) \times 100\%$  exact Blythe-Still-Casella confidence interval where  $\alpha \leq 0.05$ ; these values can be found using the StatXact software package. However, if the observed total number of biomarker-positive subjects is much different from the expected value, then the cut-offs ( $k^-$  and  $k^+$ ) can be changed using the confidence interval approach aiming to preserve the desired operating features of the design. Moreover, the value of  $k_2^+$  needed also during the second stage of the trial for decision making can be acquired using either the confidence interval approach or through the calculation of exact binomial probabilities.

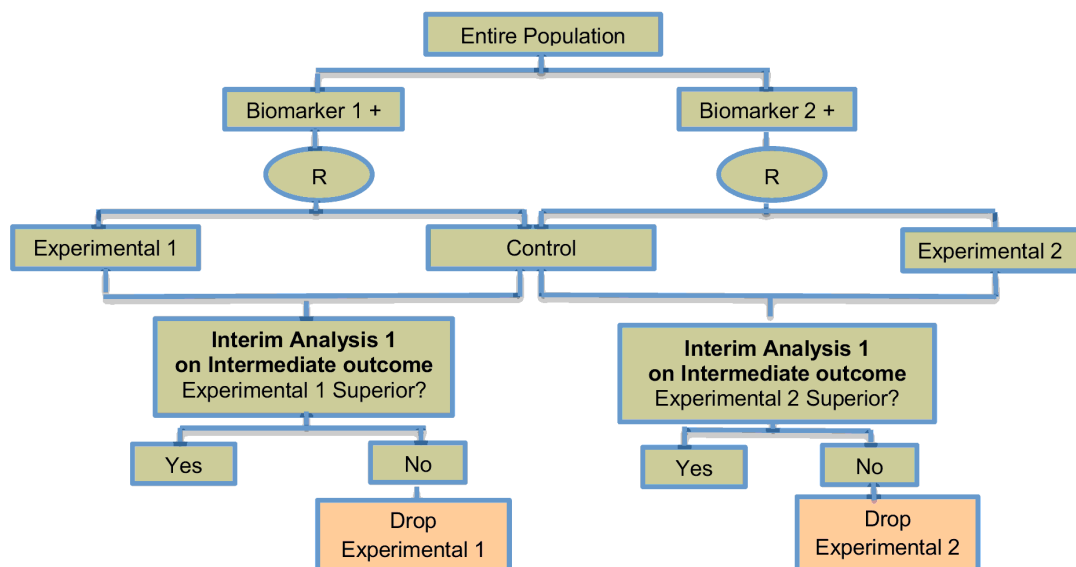
**Statistical/practical considerations:** The Adaptive Parallel Simon two-stage design may be considered as a simple approach with reasonable operating characteristics which can result in sample size savings as compared to the Simon two-stage design [88]. Similarly to Simon's two-stage design, early termination of the study is not allowed during the initial stage of the trial for efficacy in a single biomarker-defined subgroup. Additionally, this approach requires the pre-specification of appropriate response rates in both biomarker-positive and biomarker-negative subgroups which may be difficult.

Multi-arm multi-stage (MAMS) as originally proposed were not for biomarker designs. They aimed at testing multiple experimental treatment against a control treatment in the same trial. With the current advances, MAMS can be used in personalized medicine within the same context, however, they have also the ability to assess the impact of different biomarkers; the treatment arm within which a patient is included depends on their biomarker status. MAMS were found in 16 (14.9%) papers. They have the ability to simultaneously compare multiple experimental treatments with the standard treatment in order to achieve more reliable results in less time as compared with separate Phase II trials to assess each novel treatment individually. Depending on how long the actual endpoint takes to observe, the actual or an intermediate endpoint is used to identify both treatments for which there is an early sign of effectiveness and treatments that appear ineffective thus allowing the study to continue with the promising experimental arms and to stop the investigation of insufficient treatments. Generally, MAMS designs, according to Parmar et al. (2008) [97], are useful when (i) there are multiple promising treatments in phase II/III studies; (ii) there is no strong belief that a treatment will be more beneficial compared to another therapy; (iii) availability of adequate funds; (iv) there is an adequate number of patients to be enrolled and (v) there is an intermediate outcome measure correlated with the primary outcome measure. Parmar et al. (2008) [97] encouraged the use of the MAMS strategy in the field of oncology but highlighted that these designs should only be used when quick outcome assessment is possible [69]. There are two variants referred to as i) Two-stage adaptive seamless design, ii) Group Sequential design to the MAMS designs with differences occurring in its methodology. Information about these variants can be found in Appendix A.1, section “Variations of Multi-arm multi-stage (MAMS) design”. Some examples of actual trials which use the MAMS approach are the following: i) GOG-182 [20, 97, 102], ii) STAMPEDE [93, 97], iii) ICON6 [93, 97, 109], iv) FOCUS4 trial [69, 103].

**Design:** Figure 2.7 illustrates a MAMS design where the first stage of the trial (the Phase II stage) involves randomization within one of two arms which



simultaneously compare two experimental treatments with the standard of care (control) using an intermediate outcome measure (e.g. progression free survival). The arm within which a patient is included depends on their biomarker status, for example patients positive for biomarker 1 may be randomized in arm 1 to either standard of care or experimental treatment 1 whilst patients positive for biomarker 2 may be randomized in arm 2 to either standard of care or experimental treatment 2. At the end of this first stage, an interim analysis is undertaken in each arm, comparing the experimental treatment with standard of care. Depending on the outcome of the interim analysis, accrual of patients either continues within an arm to the second stage of the trial or the accrual of additional patients stops within that arm. Despite the fact that some experimental treatments cannot pass the first stage, a secondary analysis can be conducted for each of these treatment arms comparing them with the standard of care. This approach ensures that patients are randomized to the most promising treatments which were selected at the first stage of the study.



**Figure 2.7.** Multi-arm multi-stage (MAMS) design. “R” refers to randomization of patients.

**Methodology:** At the interim stage, in the case where the observed effect size in an experimental arm is greater than a predefined critical value, accrual of patients continues within that arm to the second stage of the trial until the pre-specified number of events on the primary outcome (e.g. overall survival) measure is reached, otherwise the accrual of additional patients stops within that arm and the

corresponding novel treatment does not enter the second stage of the trial. The aforementioned predefined critical value is calculated for each stage of the study in a way that takes into account whether the null hypothesis can be rejected at the level of the probability of the continuation of the study to the next stage should the null hypothesis be true as Parmar et al. (2008) [97] state.

The stopping thresholds are based on test statistics, resulting in dropping experimental arms which do not show effectiveness. The allocation to each remaining arm is fixed in MAMS trials, however, it is possible to assign more patients in the control treatment group than to the experimental arms which can yield small gains in efficiency over balanced randomization as Wason and Trippa (2014) [69] highlighted; this strategy has been used in practice with the STAMPEDE trial where the control arm is compared with five experimental treatments with the corresponding randomization ratio 2:1:1:1:1 [93]. MAMS approach could be designed with either a fixed sample size by fixing the number of patients enrolled at each stage or a fixed number of patients enrolled per arm per stage [69].

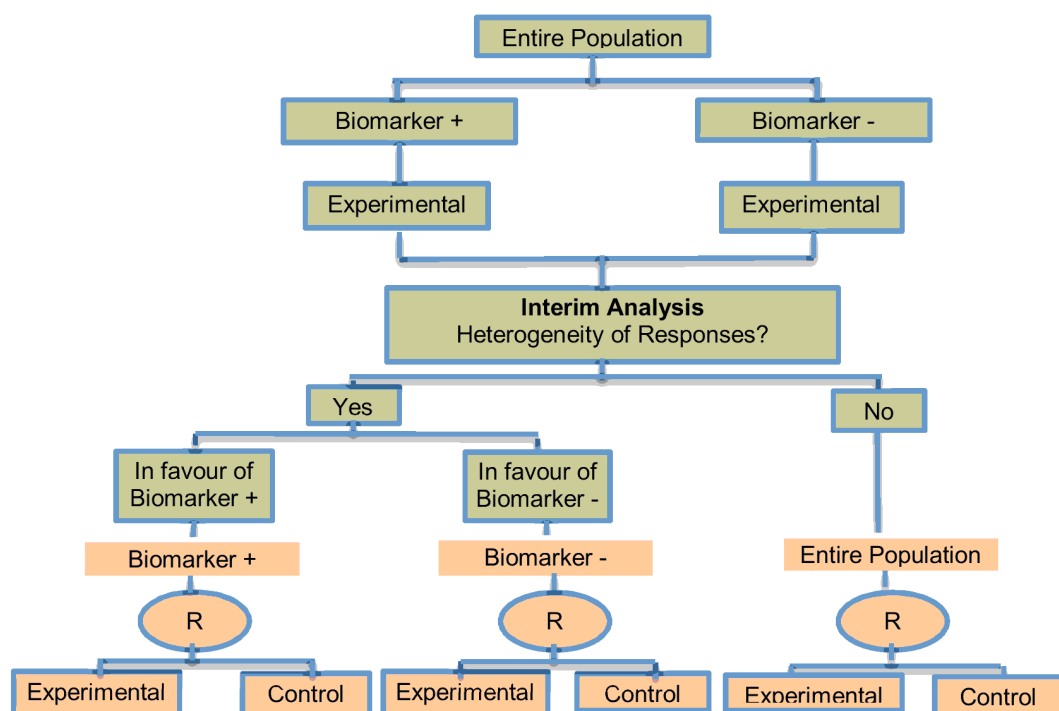
The methodology has mainly focused on situations where the primary endpoint is assumed normally distributed or time-to-event [69]. Two papers discuss MAMS designs with the normally distributed endpoint [94, 110], whilst a time-to-event endpoint is used by Royston et al. (2003) [99]. Freely-available software in Stata for calculating sample size was proposed by Barthel et al. (2009) [97] for MAMS trials [93].

A recent article by Wason et al. (2015) [111] proposed a new Bayesian adaptive design for clinical trials with biomarkers and linked treatments in multi-arm phase II trials. It is a novel approach combining the methodology used for BATTLE, I-SPY 2 and FOCUS 4 trial, which results in significant power to identify treatment effects. This novel trial design could be used for simultaneously testing several predictive biomarkers and new experimental treatments in a more cost-effective and rapid way.

**Statistical/practical considerations:** MAMS designs as compared with testing each experimental treatment in separate large-scale two-armed trials not only shorten the length of time required and reduce the costs due to the fact that they assess several experimental treatments at the same time while using a smaller number of individuals as some experimental treatment arms are dropped early. Despite the aforementioned benefits, researchers are faced with operational challenges and difficulties in building-up such designs.

#### 2.2.2.7. Stratified adaptive design

Tournoux-Facon et al. (2011) [89] proposed a new Adaptive Stratified phase II design based on the multiple-stage Fleming design [112]. A single article (0.93%) of our review referred to this approach. It is an alternative approach to dealing with stratification in a phase II setting and aims to demonstrate whether an experimental treatment (no control arm is included, thus it's about a single arm approach) is beneficial for at least one biomarker-defined subgroup rather than the entire study population.



**Figure 2.8.** Stratified adaptive design. “R” refers to randomization of patients.

**Methodology:** Decision making and the number of patients used at the second stage of the trial are based on the observed response rates during the first stage of the trial. This approach depends on the identification of heterogeneity between the two biomarker-defined subgroups (positive and negative subgroups). Heterogeneity is identified when the observed response rate in one of the biomarker-defined subgroups is less than  $\pi_{0i}$  (defined as the probability of response in one of the biomarker-defined subsets below which the novel treatment is considered to be a low-activity treatment, where  $i$  denotes each biomarker-defined subgroup; the value of 0.25 is used for the  $\pi_{0i}$  by Tournoux-Facon et al. (2011) [89]), whereas the other subset has a response rate greater than  $\pi_{0i}$ . The subset for which the observed response rate is less than  $\pi_{0i}$  is considered clinically insignificant, and therefore cannot continue to the second stage of the trial. Only the subgroup with response rate greater than  $\pi_{0i}$  therefore enters the second stage where the study can continue as a randomized Phase III trial comparing the novel treatment which has proved to be effective with the standard of care. More precisely, the identification of heterogeneity of responses is performed by calculating the symmetric interval of probability around  $\pi_{0i}$  at each stage (only a symmetric interval is observed due to binomial calculation). When the first stage of the design is terminated, in case that the cumulative number of responses for one of the biomarker-defined subset is less than/greater than the lower/upper boundary of the aforementioned symmetric interval of probability and the cumulative number of responses for the other biomarker-defined subgroup is greater than/less than the upper/lower boundary of the symmetric interval, then the responses between the two subsets are considered heterogeneous; otherwise, the treatment effect is similar in the two subsets, consequently, the trial continues without selecting any biomarker-defined subset. After the identification of heterogeneity of responses, conclusions at the end of the first stage of the trial are made according to decision rules based on specific thresholds which are determined by iterations using a Fleming two-stage approach [112]; a single-arm design which permits early termination of the trial for either efficacy or inefficacy of the treatment.

The adaptive stratified design has a number of differences from the Adaptive Parallel Simon two-stage design proposed by Jones and Holmgren (2007) [85] and the

global one-sample test for response rates for stratified phase II clinical trials proposed by London and Chang (2005) [113]. First and foremost, the adaptive stratified design permits futility or efficacy of the study as it is a strategy based on a Fleming design [112]. On the contrary, the two aforementioned methods are based on the Simon design and do not make the discontinuation for efficacy or futility of the study possible. Additionally, the stratification approach used in the design provided by Tournoux-Facon et al. (2011) [89] is utilized in order to target the patients who are most likely to respond to a novel treatment, whereas, stratification in the design by London and Chang (2005) [113] aims to ameliorate the power of the overall test.

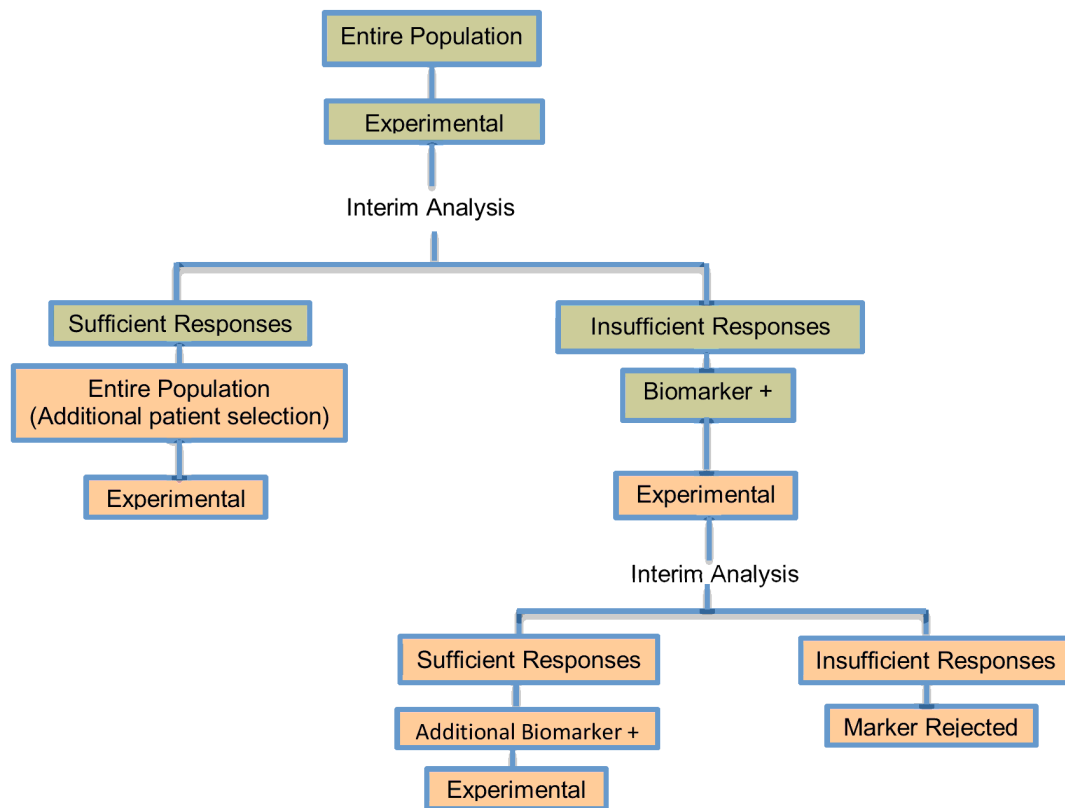
**Statistical/practical considerations:** Tournoux-Facon et al. (2011) [89] state several benefits, such as the possibility of early termination for efficacy or inefficacy of the novel treatment according to the results of the interim analysis (first stage). Moreover, this approach can be considered more ethical due to the fact that it identifies a particular biomarker-defined subpopulation for which the novel treatment can be effective and thus avoids conducting a study with patients exposed to toxic treatments. Additionally, this strategy ameliorates targeting of the populations entering phase III trials. No statistical challenges have been identified for this type of trial design so far.

#### *2.2.2.8. Tandem two stage design*

---

The tandem two-stage design was discussed in 5 (4.7%) papers. It was proposed by Pusztaï et al. (2007) [104] and it is composed of 2 optimal trials in a Phase II setting (Figure 2.9). This design was proposed for rapid biomarker assessment in settings where we don't know the activity of a novel treatment in the unselected population but there is at least one candidate predictor of response. This approach can identify whether the novel treatment is effective in the unselected patients, and if it is not, can tell us if the predictor can enrich the responding population [104]. Only an experimental treatment arm is included in this design and not a control treatment arm, thus this approach can be considered a single-arm approach. An example of

actual trial which uses the tandem two stage approach is the NCT00735917 [90, 92, 114, 115].



**Figure 2.9.** Tandem two stage design. “R” refers to randomization of patients.

**Design:** In this design, a predefined biomarker is assumed. In the first stage of the trial, patients from the entire population enter the trial irrespective of their biomarker status. An interim analysis is then undertaken and if a sufficient number of events (defined in terms of clinical benefit rate or response rate) have been observed during the first stage, the study proceeds to a second stage whereby further patients are accrued from the unselected population to establish the benefit rate more precisely in unselected patients. However, if an insufficient number of events have been observed during the first stage, rather than stopping accrual for futility, a second trial commences whereby its first stage involves continued accrual of biomarker-positive patients only. An interim analysis is then conducted and if a sufficient number of events have been occurred, this second trial continues into a second stage of biomarker-marker positive patient accrual. Otherwise, if an insufficient number of events have occurred, the predefined biomarker is rejected.

**Methodology:** A second phase in the trial design is considered due to the fact that the small number of individuals used in the first phase of the study (typically  $n_1 < 25$ ) is likely to include insufficient number of biomarker-positive individuals in order to decide whether the novel treatment benefits this particular biomarker-defined subset. In terms of defining what constitutes a ‘sufficient number’, Puzstai et al. (2007) [104] suggest the use of a non-informative prior distribution for clinical benefit rate of  $\beta(1, 1)$  and make recommendations for the early stopping rules. More precisely, Puzstai et al. (2007) [104], given a certain value for the targeted level of activity of the novel treatment (i.e. 25% clinical benefit rate), suggest that the trial should stop early for futility if the conditional power (i.e. the chance to reach the aforementioned targeted level of activity) is equal or less than 7.5% in the following cases: (i) at the first 9 evaluated patients there is no one who responds to treatment; (ii) at the first 15 evaluated patients there is only one individual who responds to treatment and (iii) at the first 20 evaluated patients there are only 2 individuals who respond to treatment.

The sample size for this approach is calculated with the same rules as a classic two-stage or Bayesian phase II design [104] where criteria for specifying the sample size are used (e.g. one criterion is to choose a sample size so that if there is no early termination of the trial and the trial accrues the entire population the posterior of the experimental treatment success rate reaches a specified degree of precision). The sample size calculations are discussed in two papers [116, 117].

**Statistical considerations:** The two trials within this design could be conducted separately, as two independent trials for the unselected individuals and for the biomarker-positive individuals, however, this can result in larger duration and costs, therefore it would be better to run the two trials as a single study (see Table 2.1 for further details). Additionally, this approach enables the estimation of response rates in both biomarker-negative and biomarker-positive patients.

## 2.3. Discussion

---

The review has demonstrated ambiguity and confusion regarding biomarker-guided adaptive designs proposed by different authors. For example, different authors described the same trial design by naming it differently (see Table 2.1). In this review, we focus on 8 types of such designs. There are several reasons why these design strategies are becoming an appealing approach to a great extent. The main reason is their application to real clinical practice and their ability to evaluate both multiple experimental treatments and biomarkers simultaneously. Hence, multiple questions can be answered just in a single trial [48]. During the progression of the trial alterations are permitted, and consequently, any potential incorrect hypothesis made at the beginning of the trial can be modified. Many authors note that these strategies are ethical in terms of safety and efficacy as they attempt to tailor the appropriate treatment to the right population at the right time [10, 33, 37, 40, 46, 55, 70, 118, 119]. The required number of patients needed for the enrollment in the trial can be modified according to the results from interim analysis (e.g. stop accrual or increase sample size) and the duration of the trial can be minimized as they allow for dropping early treatments which show poor performance. Also, due to alterations, e.g. if incrementation of the sample size is suggested as the study progresses, higher power to demonstrate a treatment effect may be achieved [120]. Furthermore, it has been argued that during the adaptation process, preservation of type I and type II error rates may be attained through the appropriate choice of statistical parameters [26].

Despite the aforementioned advantages, there are a considerable number of challenges which should be carefully investigated before making a decision. Their implementation may be considered a poor choice when there is already a high quality retrospective dataset available which includes information both on biomarker status and on long-term follow-up, since in such a situation an analysis of this dataset to identify a biomarker subgroup would likely be more efficient as a first stage as opposed to incorporating this first stage into the trial itself [120]. Also, they can be complex in terms of logistic issues such as maintaining trial integrity, minimizing



operational bias [33, 45, 48, 52] and the involved perspectives of regulatory agencies (e.g. what level of adaptation will be acceptable to the regulatory agencies) [121]. In addition, adaptations, of which statistical validity may be challenging, can lead to notable modifications yielding a complicated trial totally different from the initial study [33, 37, 40]. Consequently, it could diverge from the original question which researchers expect to answer. Furthermore, statistical validity of conclusions can be influenced to a great extent as unexpected bias or variation may be introduced during the course of the trial making the interpretation of results greatly complex [33, 122, 123]. The inserted operational bias occurred by the modifications in the trial design augments the likelihood of making a false conclusion that the treatment is beneficial to certain patients [33, 37, 40, 71, 118]. It is necessary that adaptive designs are planned in such a way that allows for controlling both Type I and Type II error rates [69]. Additionally, from a statistical viewpoint, adaptive designs based on Bayesian methods are considered computationally intensive [55] and estimations of Type I error rate can be inaccurate. Problems of statistical testing may also arise and applying the statistical methods can be very challenging without the availability of appropriate software packages to facilitate the implementation of adaptive designs (e.g. computational intensive demands of Bayesian methods) [33, 40, 45, 48, 52]. A number of obstacles and barriers related to the conduct of adaptive designs in practice in Phase III trials is addressed in a recent paper [124]; several key stakeholders in clinical trials research have been interviewed and some of the highlighted difficulties expressed during this study were the lack of appropriate knowledge and familiarity of these designs in the biostatistics community, insufficient time and funding structure, additional work required due to the complexity of such designs and the needed statistical expertise and appropriate software.

However, adaptive designs will continue to hold a prominent place in the era of personalized medicine, and hence, further developments and discussion are of utmost importance in order to enhance clinical research. In conducting such further developments and discussion, investigators should take account of the following points in particular (i) regulatory and logistical issues; (ii) statistical challenges

including the control of the false-positive rate, power of the study and treatment effect estimation; (iii) the unexpected bias likely to be introduced during the adaptation process and (iv) the potential increased cost and time. Further, the different designs proposed so far for adaptive trials need to be better understood by the research community, as the proper use of such designs can result in a great increase in the efficiency of a trial and boost the development of novel treatments. By conducting this methodological review, we contribute to the knowledge enhancement of researchers regarding the biomarker-guided adaptive trial designs.

The characteristics and methodology of the eight main designs are discussed here, whilst information on their variations are summarized in Table A.1, Appendix A.1. Additional references for these variations are provided in [125-142].

In the current chapter we presented the adaptive trial designs, whereas in the next chapter (Chapter 3) we will focus on the second broad category, the so-called non-adaptive trial designs.

## 2.4. References

---

1. George SL. Statistical issues in translational cancer research. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(19):5954-8. doi: 10.1158/1078-0432.CCR-07-4537.
2. Chabner B. Advances and challenges in the use of biomarkers in clinical trials. *Clinical advances in hematology & oncology: H&O*. 2008; 6(1):42-3.
3. Shi Q, Mandrekar SJ, Sargent DJ. Predictive biomarkers in colorectal cancer: usage, validation, and design in clinical trials. *Scandinavian journal of gastroenterology*. 2012; 47(3):356-62. doi: 10.3109/00365521.2012.640836.
4. Pihlstrom BL, Barnett ML. Design, operation, and interpretation of clinical trials. *Journal of dental research*. 2010; 89(8):759-72. doi: 10.1177/0022034510374737.

5. Rigatto C, Barrett BJ. Biomarkers and surrogates in clinical studies. *Methods in molecular biology* (Clifton, NJ). 2009; 473:137-54. doi: 10.1007/978-1-59745-385-1\_8.
6. Mandrekar SJ, An M-W, Sargent DJ. A review of phase II trial designs for initial marker validation. *Contemporary clinical trials*. 2013; 36(2):597-604. doi: 10.1016/j.cct.2013.05.001.
7. Karuri SW, Simon R. A two-stage Bayesian design for co-development of new drugs and companion diagnostics. *Statistics in medicine*. 2012; 31(10):901-14. doi: 10.1002/sim.4462.
8. Matsui S. Genomic biomarkers for personalized medicine: development and validation in clinical studies. *Computational and mathematical methods in medicine*. 2013; 2013:865980-. doi: 10.1155/2013/865980.
9. Buyse M, Michiels S. Omics-based clinical trial designs. *Current opinion in oncology*. 2013; 25(3):289-95. doi: 10.1097/CCO.0b013e32835ff2fe.
10. Wu W, Shi Q, Sargent DJ. Statistical considerations for the next generation of clinical trials. *Seminars in oncology*. 2011; 38(4):598-604. doi: 10.1053/j.seminoncol.2011.05.014.
11. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2005; 23(9):2020-7. doi: 10.1200/JCO.2005.01.112.
12. Chen JJ, Lu T-P, Chen D-T, Wang S-J. Biomarker adaptive designs in clinical trials. *Translational Cancer Research*. 2014; 3(3):279-92.
13. Freidlin B, Sun Z, Gray R, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2013; 31(25):3158-61. doi: 10.1200/JCO.2012.48.3826.

14. Gosho M, Nagashima K, Sato Y. Study designs and statistical analyses for biomarker research. *Sensors* (Basel, Switzerland). 2012; 12(7):8966-86. doi: 10.3390/s120708966.
15. Ming-Wen An SJM, Daniel JS. Biomarkers-guided targeted drugs: new clinical trials design and practice necessity. *Advances in Personalized Cancer Management*. 2011; 30-41. doi: 10.2217/ebo.11.87.
16. Buyse M. Towards validation of statistically reliable biomarkers. *European Journal of Cancer Supplements*. 2007; 5(5):89-95. doi: 10.1016/S1359-6349(07)70028-9.
17. Lee CK, Lord SJ, Coates AS, Simes RJ. Molecular biomarkers to individualise treatment: assessing the evidence. *The Medical journal of Australia*. 2009; 190(11):631-6.
18. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized medicine*. 2010; 7(1):33-47. doi: 10.2217/pme.09.49.
19. Fraser GAM, Meyer RM. Biomarkers and the design of clinical trials in cancer. *Biomarkers in medicine*. 2007; 1(3):387-97. doi: 10.2217/17520363.1.3.387.
20. Mandrekar SJ, Sargent DJ. Design of clinical trials for biomarker research in oncology. *Clinical investigation*. 2011; 1(12):1629-36. doi: 10.4155/CLI.11.152.
21. Simon R. Advances in clinical trial designs for predictive biomarker discovery and validation. *Current Breast Cancer Reports*. 2009; 1(4):216-21. doi: 10.1007/s12609-009-0030-4.
22. Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*. 2013; 105(22):1677-83. doi: 10.1093/jnci/djt282.

23. Bradley E. Incorporating biomarkers into clinical trial designs: points to consider. *Nature biotechnology*. 2012; 30(7):596-9. doi: 10.1038/nbt.2296.
24. Beckman RA, Clark J, Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature reviews Drug discovery*. 2011; 10(10):735-48. doi: 10.1038/nrd3550.
25. Young KY, Laird A, Zhou XH. The efficiency of clinical trial designs for predictive biomarker validation. *Clinical trials (London, England)*. 2010; 7(5):557-66. doi: 10.1177/1740774510370497.
26. Lee JJ, Xuemin G, Suyu L. Bayesian adaptive randomization designs for targeted agent development. *Clinical trials (London, England)*. 2010; 7(5):584-96. doi: 10.1177/1740774510373120.
27. Simon R. Clinical trials for predictive medicine: new challenges and paradigms. *Clinical trials (London, England)*. 2010; 7(5):516-24. doi: 10.1177/1740774510366454.
28. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points--the challenge of statistical validation. *Nature reviews Clinical oncology*. 2010; 7(6):309-17. doi: 10.1038/nrclinonc.2010.43.
29. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2009; 27(24):4027-34. doi: 10.1200/JCO.2009.22.3701.
30. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal of biopharmaceutical statistics*. 2009; 19(3):530-42. doi: 10.1080/10543400902802458.
31. Hoering A, Leblanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clinical cancer research: an official journal of the American*

Association for Cancer Research. 2008; 14(14):4358-67. doi: 10.1158/1078-0432.CCR-08-0288.

32. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. *Nature reviews Drug discovery*. 2012; 11(3):201-14. doi: 10.1038/nrd3651.

33. Chow S-C. Adaptive clinical trial design. *Annual review of medicine*. 2014; 65:405-15. doi: 10.1146/annurev-med-092012-112310.

34. Chow S-C, Chang M, Pong A. Statistical consideration of adaptive methods in clinical development. *Journal of biopharmaceutical statistics*. 2005; 15(4):575-91. doi: 10.1081/BIP-200062277.

35. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development--an Executive Summary of the PhRMA Working Group. *Journal of biopharmaceutical statistics*. 2006; 16(3):275-83; discussion 85-91, 93-8, 311-2. doi: 10.1080/10543400600614742.

36. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*. 2009; 28(10):1445-63. doi: 10.1002/sim.3559.

37. Chow S-C, Tu Y-H. On Two-stage Seamless Adaptive Design in Clinical Trials. *Journal of the Formosan Medical Association = Taiwan yi zhi*. 2008; 107(12 Suppl):52-60.

38. Maharaj R. Vasopressors and the search for the optimal trial design. *Contemporary clinical trials*. 2011; 32(6):924-30. doi: 10.1016/j.cct.2011.07.010.

39. Vandemeulebroecke M. Group sequential and adaptive designs - a review of basic concepts and points of discussion. *Biometrical journal Biometrische Zeitschrift*. 2008; 50(4):541-57. doi: 10.1002/bimj.200710436.

40. Chow S-C, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet journal of rare diseases*. 2008; 3:11-. doi: 10.1186/1750-1172-3-11.
41. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. *Pharmaceutical statistics*. 2007; 6(3):205-16. doi: 10.1002/pst.302.
42. Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities. *Trials*. 2012; 13:145-. doi: 10.1186/1745-6215-13-145.
43. Ananthakrishnan R, Menon S. Design of oncology clinical trials: a review. *Critical reviews in oncology/hematology*. 2013; 88(1):144-53. doi: 10.1016/j.critrevonc.2013.03.007.
44. Orloff JJ, Stanski D. Innovative approaches to clinical development and trial design. *Annali dell'Istituto superiore di sanità*. 2011; 47(1):8-13. doi: 10.4415/ANN\_11\_01\_03.
45. Dragalin V. An introduction to adaptive designs and adaptation in CNS trials. *European neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology*. 2011; 21(2):153-8. doi: 10.1016/j.euroneuro.2010.09.004.
46. Coffey CS, Kairalla JA. Adaptive clinical trials: progress and challenges. *Drugs in R&D*. 2008; 9(4):229-42.
47. Freidlin B, Korn EL. Biomarker-adaptive clinical trial designs. *Pharmacogenomics*. 2010; 11(12):1679-82. doi: 10.2217/pgs.10.153.
48. Heckman-Stoddard BM, Smith JJ. Precision medicine clinical trials: defining new treatment strategies. *Seminars in oncology nursing*. 2014; 30(2):109-16. doi: 10.1016/j.soncn.2014.03.004.
49. Galanis E, Wu W, Sarkaria J, Chang SM, Colman H, Sargent D, et al. Incorporation of biomarker assessment in novel clinical trial designs: personalizing brain tumor

treatments. *Current oncology reports*. 2011; 13(1):42-9. doi: 10.1007/s11912-010-0144-x.

50. An M-W, Mandrekar SJ, Sargent DJ. A 2-stage phase II design with direct assignment option in stage II for initial marker validation. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2012; 18(16):4225-33. doi: 10.1158/1078-0432.CCR-12-0686.

51. Van Schaeybroeck S, Allen WL, Turkington RC, Johnston PG. Implementing prognostic and predictive biomarkers in CRC clinical trials. *Nature reviews Clinical oncology*. 2011; 8(4):222-32. doi: 10.1038/nrclinonc.2011.15.

52. Ang M-K, Tan S-B, Lim W-T. Phase II clinical trials in oncology: are we hitting the target? Expert review of anticancer therapy. 2010; 10(3):427-38. doi: 10.1586/era.09.178.

53. Eickhoff JC, Kim K, Beach J, Kolesar JM, Gee JR. A Bayesian adaptive design with biomarkers for targeted therapies. *Clinical trials (London, England)*. 2010; 7(5):546-56. doi: 10.1177/1740774510372657.

54. Berry DA. Adaptive clinical trials in oncology. *Nature reviews Clinical oncology*. 2012; 9(4):199-207. doi: 10.1038/nrclinonc.2011.165.

55. Lee JJ, Chu CT. Bayesian clinical trials in action. *Statistics in medicine*. 2012; 31(25):2955-72. doi: 10.1002/sim.5404.

56. Berry DA. Bayesian clinical trials. *Nature reviews Drug discovery*. 2006; 5(1):27-36. doi: 10.1038/nrd1927.

57. Simon R. Clinical trials for predictive medicine. *Statistics in medicine*. 2012; 31(25):3031-40. doi: 10.1002/sim.5401.



58. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature reviews Clinical oncology*. 2014; 11(2):81-90. doi: 10.1038/nrclinonc.2013.218.
59. Scher HI, Nasso SF, Rubin EH, Simon R. Adaptive clinical trial designs for simultaneous testing of matched diagnostics and therapeutics. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2011; 17(21):6634-40. doi: 10.1158/1078-0432.CCR-11-1105.
60. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2010; 16(2):691-8. doi: 10.1158/1078-0432.CCR-09-1357.
61. Coyle VM, Johnston PG. Genomic markers for decision making: what is preventing us from using markers? *Nature reviews Clinical oncology*. 2010; 7(2):90-7. doi: 10.1038/nrclinonc.2009.214.
62. Berry DA, Herbst RS, Rubin EH. Reports from the 2010 Clinical and Translational Cancer Research Think Tank meeting: design strategies for personalized therapy trials. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2012; 18(3):638-44. doi: 10.1158/1078-0432.CCR-11-2018.
63. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2013; 19(17):4578-88. doi: 10.1158/1078-0432.CCR-12-3722.
64. Baker SG, Kramer BS, Sargent DJ, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. *Discovery medicine*. 2012; 13(70):187-92.
65. Di Maio M, Gallo C, De Maio E, Morabito A, Piccirillo MC, Gridelli C, et al. Methodological aspects of lung cancer clinical trials in the era of targeted agents.

Lung cancer (Amsterdam, Netherlands). 2010; 67(2):127-35. doi: 10.1016/j.lungcan.2009.10.001.

66. Simon R. The use of genomics in clinical trial design. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(19):5984-93. doi: 10.1158/1078-0432.CCR-07-4531.

67. Simon R. Development and validation of biomarker classifiers for treatment selection. *Journal of Statistical Planning and Inference*. 2008; 138:308-20. doi: 10.1016/j.jspi.2007.06.010. PubMed PMID: S037837580700242X.

68. Simon R. Biomarker based clinical trial design. *Chinese clinical oncology*. 2014; 3(3):39. doi: 10.3978/j.issn.2304-3865.2014.02.03.

69. Wason JMS, Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in medicine*. 2014; 33(13):2206-21. doi: 10.1002/sim.6086.

70. Sato Y, Laird NM, Yoshida T. Biostatistic tools in pharmacogenomics - advances, challenges, potential. *Current pharmaceutical design*. 2010; 16(20):2232-40.

71. Korn EL, Freidlin B. Outcome - adaptive randomization: is it useful? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2011; (6):771-6. doi: 10.1200/JCO.2010.31.1423.

72. Lai TL, Lavori PW, Shih M-CI, Sikic BI. Clinical trial designs for testing biomarker-based personalized therapies. *Clinical trials (London, England)*. 2012; 9(2):141-54. doi: 10.1177/1740774512437252.

73. Gold KA, Kim ES, Lee JJ, Wistuba II, Farhangfar CJ, Hong WK. The BATTLE to personalize lung cancer prevention through reverse migration. *Cancer prevention research (Philadelphia, Pa)*. 2011; 4(7):962-72. doi: 10.1158/1940-6207.CAPR-11-0232.

74. Lai TL, Liao OY-W, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemporary clinical trials*. 2013; 36(2):651-63. doi: 10.1016/j.cct.2013.08.007.
75. Younes A, Berry DA. From drug discovery to biomarker-driven clinical trials in lymphoma. *Nature reviews Clinical oncology*. 2012; 9(11):643-53. doi: 10.1038/nrclinonc.2012.156.
76. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert review of molecular diagnostics*. 2011; 11(2):171-82. doi: 10.1586/erm.10.120.
77. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer--a step toward personalized medicine. *Clinical trials (London, England)*. 2008; 5(3):181-93. doi: 10.1177/1740774508091815.
78. European Medicines Agency. Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection London; 2011 [updated [cited 2012 Jul 3]; cited 2015 10 Oct]. Available from: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2011/07/WC500108672.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/07/WC500108672.pdf).
79. Liu A, Liu C, Li Q, Yu KF, Yuan VW. A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Clinical trials (London, England)*. 2010; 7(5):537-45. doi: 10.1177/1740774510378695.
80. Wang S-J, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical statistics*. 2007; 6(3):227-44. doi: 10.1002/pst.300.

81. Wang S-J. Biomarker as a classifier in pharmacogenomics clinical trials: a tribute to 30th anniversary of PSI. *Pharmaceutical statistics*. 2007; 6(4):283-96. doi: 10.1002/pst.316.
82. Emerson SS, Fleming TR. Adaptive methods: telling "the rest of the story". *Journal of biopharmaceutical statistics*. 2010; 20(6):1150-65. doi: 10.1080/10543406.2010.514457.
83. Wang S-J, Hung HMJ, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biometrical journal Biometrische Zeitschrift*. 2009; 51(2):358-74. doi: 10.1002/bimj.200900003.
84. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert opinion on medical diagnostics*. 2008; 2(6):721-9. doi: 10.1517/17530059.2.6.721.
85. Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. *Contemporary clinical trials*. 2007; 28(5):654-61. doi: 10.1016/j.cct.2007.02.008.
86. Wang S-J. Adaptive strategy versus adaptive design in pharmacogenomics or pharmacogenetics clinical trials. *Journal of the Formosan Medical Association*. 2008; 107(S18–S26).
87. Ho TW, Pearlman E, Lewis D, Hämäläinen M, Connor K, Michelson D, et al. Efficacy and tolerability of rizatriptan in pediatric migraineurs: results from a randomized, double-blind, placebo-controlled trial using a novel adaptive enrichment design. *Cephalalgia: an international journal of headache*. 2012; 32(10):750-65. doi: 10.1177/0333102412451358.
88. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled clinical trials*. 1989; 10(1):1-10.

89. Tournoux-Facon C, De Rycke Y, Tubert-Bitter P. Targeting population entering phase III trials: a new stratified adaptive phase II design. *Statistics in medicine*. 2011; 30(8):801-11. doi: 10.1002/sim.4148.
90. McShane LM, Hunsberger S, Adjei AA. Effective incorporation of biomarkers into phase II trials. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2009; 15(6):1898-905. doi: 10.1158/1078-0432.CCR-08-2033.
91. Simon R, Polley E. Clinical trials for precision oncology using next-generation sequencing. *Personalized Medicine*. 2013; 10:485-95. doi: 10.2217/pme.13.36.
92. Andre F. Study CTKI258A2202: A multicenter, open-label phase II trial of dovitinib (TKI258) in FGFR1-amplified and nonamplified HER2-negative metastatic breast cancer: ASCO; 2010 [cited 2015 10 Oct]. Available online: <http://meetinglibrary.asco.org/content/52807-74>.
93. Sydes MR, Parmar MKB, James ND, Clarke NW, Dearnaley DP, Mason MD, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials*. 2009; 10:39. doi: 10.1186/1745-6215-10-39.
94. Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in medicine*. 2012; 31(30):4269-79. doi: 10.1002/sim.5513.
95. Ferraldeschi R, Attard G, de Bono JS. Novel strategies to test biological hypotheses in early drug development for advanced prostate cancer. *Clinical chemistry*. 2013; 59(1):75-84. doi: 10.1373/clinchem.2012.185157.
96. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. *Clinical Trials*. 2010; 7(5):567-73.
97. Parmar MKB, Barthel FMS, Sydes M, Langley R, Kaplan R, Eisenhauer E, et al. Speeding up the evaluation of new agents in cancer. *Journal of the National Cancer Institute*. 2008; 100(17):1204-14. doi: 10.1093/jnci/djn267.

98. Barthel FMS, Parmar MKB, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design--a reanalysis of 4 trials. *Trials*. 2009; 10:21-. doi: 10.1186/1745-6215-10-21.
99. Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in medicine*. 2003; 22(14):2239-56. doi: 10.1002/sim.1430.
100. Chow S-C, Chang M. *Adaptive Design Methods in Clinical Trials*, Second Edition Boca Raton, FL: Chapman & Hall/CRC Press Biostatistics Series; 2011.
101. Barthel FMS, Royston P, Parmar MKB. A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome. *Stata Journal*. 2009; 9(4):505–23.
102. Copeland LJ, Bookman M, Trimble E. Clinical trials of newer regimens for treating ovarian cancer: the rationale for Gynecologic Oncology Group Protocol GOG 182-ICON5. *Gynecologic oncology*. 2003; 90(2 Pt 2):S1-7.
103. Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, et al. Evaluating many treatments and biomarkers in oncology: a new design. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2013; 31(36):4562-8. doi: 10.1200/JCO.2013.50.7905.
104. Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2007; 13(20):6080-6. doi: 10.1158/1078-0432.CCR-07-0809.
105. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2005; 11(21):7872-8. doi: 10.1158/1078-0432.CCR-05-0605.

106. Moyé LA, Deswal A. Trials within Trials. *Controlled Clinical Trials*. 2001; 22(6):605.
107. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75(4):800.
108. Novartis Pharmaceuticals. A Multi-center, Open Label Phase II Trial of TKI258 in FGFR1 Amplified and Non-amplified Metastatic or Advanced HER2 Negative Breast Cancer: ClinicalTrials.gov; 2009 [cited 2015 10 Oct]. Available online: <https://clinicaltrials.gov/ct2/show/NCT00958971?term=NCT00958971&rank=1>.
109. Medical Research Council. A Randomised, Placebo-controlled, Trial of Concurrent Cediranib [AZD2171] (With Platinum-based Chemotherapy) and Concurrent and Maintenance Cediranib in Women With Platinum-sensitive Relapsed Ovarian Cancer: ClinicalTrials.gov; 2007 [cited 2015 10 Oct]. Available online: <https://clinicaltrials.gov/ct2/show/study/NCT00544973?term=icon6&rank=1>.
110. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*. 2012; 99(2):494-501. doi: 10.1093/biomet/ass002.
111. Wason JMS, Abraham JE, Baird RD, Gournaris I, Vallier A-L, Brenton JD, et al. A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer*. 2015.
112. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982; 38(1):143-51.
113. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine*. 2005; 24(17):2597-611.
114. Nallapareddy S., Arcaroli J., Touban B., Tan A., Foster N. R., Erlichman C., et al. A Phase II trial of saracatinib (AZD0530), an oral Src inhibitor, in previously treated

metastatic pancreatic cancer.: ASCO; 2010 [cited 2015 10 Oct]. Available from: <http://meetinglibrary.asco.org/content/1452-72>.

115. National Cancer Institute. A Phase II Trial of AZD0530 in Previously Treated Metastatic Pancreas Cancer: ClinicalTrials.gov; 2008 [cited 2015 10 Oct]. Available online: <https://clinicaltrials.gov/ct2/show/NCT00735917?term=NCT00735917&rank=1>.

116. Simon R. Cancer. Principles & practice of oncology (6th edition). Philadelphia: Lippincott Williams and Wilkins; 2001. 521-38 p.

117. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Controlled Clinical Trials*. 1994; 15(6):463-81.

118. Hung HMJ, Wang S-J, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of biopharmaceutical statistics*. 2007; 17(6):1201-10. doi: 10.1080/10543400701645405.

119. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of biopharmaceutical statistics*. 2007; 17(6):1135-61. doi: 10.1080/10543400701645215.

120. Wason J, Marshall A, Dunn J, Stein RC, Stallard N. Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *British journal of cancer*. 2014; 110(8):1950-7. doi: 10.1038/bjc.2014.156.

121. Mahajan R, Gupta K, Mahajan R, Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect. *Indian Journal of Pharmacology*. 2010; 42(4):201.

122. DeMets DL. Current development in clinical trials: issues old and new. *Statistics in medicine*. 2012; 31(25):2944-54. doi: 10.1002/sim.5405.



123. Emerson SC, Rudser KD, Emerson SS. Exploring the benefits of adaptive sequential designs in time-to-event endpoint settings. *Statistics in Medicine*. 2011; 30(11):1199-217.
124. Dimairo M, Boote J, Julious SA, Nicholl JP, Todd S. Missing steps in a staircase: a qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials*. 2015; 16(1):430.
125. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*. 2007; 99(13):1036-43. doi: 10.1093/jnci/djm022.
126. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical pharmacology and therapeutics*. 2009; 86(1):97-100. doi: 10.1038/clpt.2009.68.
127. QuantumLeap Healthcare Collaborative. I-SPY 2 Trial (Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging And moLecular Analysis 2): ClinicalTrials.gov; 2009 [cited 2015 10 Oct]. Available online: <https://clinicaltrials.gov/ct2/show/NCT01042379>.
128. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986; 73(3):751-754. doi: 10.1093/biomet/73.3.751.
129. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*. 2011; 10(4):347-56. doi: 10.1002/pst.472.
130. Mehta C, Schäfer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*. 2014; 33(26):4515-31. doi: 10.1002/sim.6272.
131. Ellenberg SS, Eisenberger MA. An efficient design for phase III studies of combination chemotherapies. *Cancer treatment reports*. 1985; 69(10):1147-54.

132. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*. 2002; 58(4):823-31.
133. Lin J-A, He P. Reinventing clinical trials: a review of innovative biomarker trial designs in cancer therapies. *British medical bulletin*. 2015; 114(1):17-27. doi: 10.1093/bmb/ldv011.
134. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics* (Oxford, England). 2013; 14(4):613-25. doi: 10.1093/biostatistics/kxt010.
135. Alexander BM, Wen PY, Trippa L, Reardon DA, Yung W-KA, Parmigiani G, et al. Biomarker-based adaptive trials for patients with glioblastoma--lessons from I-SPY 2. *Neuro-oncology*. 2013; 15(8):972-8. doi: 10.1093/neuonc/not088.
136. Freidlin B, Korn EL, Gray R. Marker Sequential Test (MaST) design. *Clinical trials* (London, England). 2014; 11(1):19-27. doi: 10.1177/1740774513503739.
137. Freidlin B, McShane LM, Polley M-YC, Korn EL. Randomized phase II trial designs with biomarkers. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2012; 30(26):3304-9. doi: 10.1200/JCO.2012.43.3946.
138. Wang S-J. Utility of adaptive strategy and adaptive design for biomarker-facilitated patient selection in pharmacogenomic or pharmacogenetic clinical development program. *Journal of the Formosan Medical Association = Taiwan yi zhi*. 2008; 107(12 Suppl):19-27.
139. Wang S-J, Hung HMJ, O'Neill R. Adaptive design clinical trials and trial logistics models in CNS drug development. *European neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology*. 2011; 21(2):159-66. doi: 10.1016/j.euroneuro.2010.09.003.
140. Tudur Smith C, Williamson PR, Beresford MW. Methodology of clinical trials for rare diseases. *Best practice & research Clinical rheumatology*. 2014; 28(2):247-62. doi: 10.1016/j.berh.2014.03.004.

141. Chow S-C, Lu Q, Tse S-K. Statistical analysis for two-stage adaptive design with different study points. *Journal of biopharmaceutical statistics*. 2007; 17(6):1163-76. doi: 10.1080/10543400701645249.

142. U. S. Food and Drug Administration. Draft Guidance for Industry—Adaptive Design Clinical Trials for Drugs and Biologics. : U.S. Food Drug Admin, Rockville, MD; 2010 [cited 2015 10 Oct]. Available online:  
<http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf>.

## **Chapter 3. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review**

---

### **3.1. Introduction**

---

To complement the previous chapter (Chapter 2) which gives detailed information regarding the biomarker-guided adaptive designs, here we report on the second broad category, the so-called non-adaptive designs, in order to provide researchers with much needed information to an extent that has not been previously available. The current chapter is based on our published paper by Antoniou et al. (2017) (see list of Publications).

The rapidly developing field of ‘personalized medicine’ [1], also known as ‘individualized medicine’, ‘stratified medicine’, or ‘precision medicine’ is allowing scientists to treat patients by providing them with a specific regimen according to their individual demographic, genomic or biological characteristics. The latter two aforementioned characteristics are collectively known as biomarkers [2]. The terms ‘personalized medicine’ and ‘individualized medicine’ often create confusion in literature, as in reality, the objective of this approach is to identify demographic- or biomarker-defined subgroups. Thus, as it still remains a population and not an individualized approach, the terms ‘stratified’ or ‘precision’ medicine are often considered to be more accurate. The National Institutes of Health Biomarkers Definitions Working Group [3] defined a biomarker to be “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [1, 4-7]. Biomarkers related to clinical outcome which are measured before treatment commences can be classified as either prognostic or predictive biomarkers. Prognostic biomarkers provide information regarding the likely progression of a disease without taking into account any specific treatment, whilst predictive

biomarkers provide information about the patient's outcome given a certain treatment, i.e. their likely response to the treatment [4, 7-34]. Prior to utilizing a patient's biomarker information in clinical practice, it is necessary that they have been robustly tested in terms of analytical validity (the results of testing a specific biomarker or biomarkers can be trusted), clinical validity (the results obtained from the test correlates with important clinical information) and clinical utility (the test will be useful in ameliorating patients' health) [9, 13, 19, 25].

A number of phase II and phase III trial designs have been proposed for testing the clinical utility of prognostic biomarkers. Due to the large amount of literature in this field, we have split our review into two broad categories, i.e. the biomarker-guided non-adaptive trial designs which are presented in the current study and the biomarker-guided adaptive trial designs. The latter are extensively discussed in Chapter 2 [35].

In this review we aim to communicate the different non-adaptive biomarker-guided trial designs, which can be either randomized or non-randomized designs (e.g., single-arm designs), proposed in the literature so far and to report on the potential advantages and weaknesses of each. Some of them include an adaptive element (see section 3.2.3), although non-adaptive in the traditional sense. Therefore, they were included in the current chapter instead of Chapter 2 [35] which describes and discusses adaptive designs.,

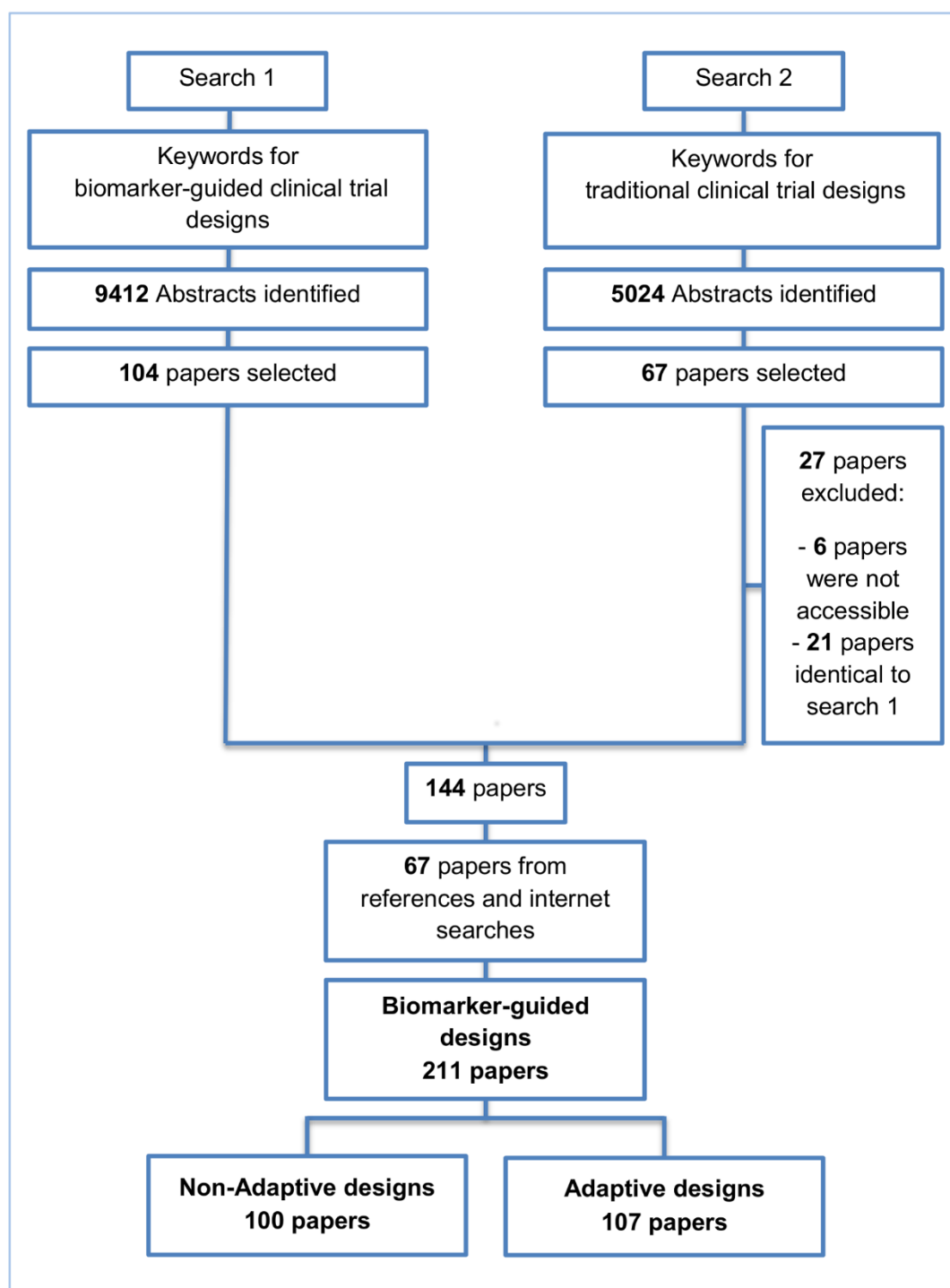
## 3.2. Methods and Findings

---

We undertook a search of the MEDLINE (Ovid) database, restricted to published papers in the English language within the previous ten years aiming to identify articles which describe and discuss biomarker-guided trial designs. Traditional trial designs, i.e. designs which do not incorporate biomarkers aiming to aid in making treatment decisions (we will refer to as 'traditional' trial designs) are part of our literature review search strategy in order to help us identify and distinguish any potential reference to biomarker-guided designs, as the finding of the

appropriate keywords in Medline database for biomarker-guided designs was challenging. Furthermore, the restriction of published papers within the past decade was made not only because of the large amount of literature in this field, but also for the identification of the most recent trial designs. Two separate strategies as illustrated in Figure 3.1 were used to identify relevant articles, and the keywords utilized in the search are presented in Appendix A.2. Our initial search resulted in 9412 and 5024 relevant titles for biomarker-guided clinical trial designs and traditional trial designs, respectively. From the 9412 papers, 104 articles were included based on their title and abstract. From the 5024 papers, 40 articles were included based on their title and abstract and after removing inaccessible articles or those already identified in the search for biomarker-guided trial designs. An additional 67 eligible papers were identified from searching both the reference list of included articles and the internet (the internet searches were performed using the same keywords as those for the Ovid strategy), making a total of 211 included papers. Of these 211 included papers, biomarker-guided non-adaptive trial designs were referred to in 100 papers; 107 papers for biomarker-guided adaptive trial designs were reviewed in our published paper Antoniou et al., 2016 [35]. In the total number of 211 papers, some papers are referred to both adaptive and non-adaptive designs. Articles from references and internet searches which did not provide further information on each broad category of biomarker-guided designs were not included. Cited books, web pages for actual trials and papers published before 2005 are also not included in these numbers. For each included paper, the following details were extracted: definition of the trial design(s) referred to in the paper, how patients were screened and/or randomized based on their biomarker status, treatment groups randomized to, as well as other key information relating to the trial design and methodology, including advantages and limitations. Where reference was made in the included papers to an actual trial which had adopted a particular biomarker-guided non-adaptive trial design, the clinical field with which the trial was associated was also recorded. However, a review of all implementations of the different trial designs in practice is beyond the scope of this chapter, and is an area for potential future work. Therefore, it is important to highlight that even where no evidence of

the implementation of a particular design was found in the papers included in our review, the design may well be currently in use in ongoing trials.



**Figure 3.1.** Flow diagram of the review process. From our search strategy a total number of 211 papers have been identified giving information regarding not only the biomarker-guided designs but also general information about personalized medicine and biomarkers. Before arriving at 211 papers, books, web pages for actual trials and papers published before 2005 were excluded. The 211 papers are split

into two overlapping sets of 100 and 107 papers. The total of 207 is less than 211 due to overlap of papers, and also due to the fact that some articles referring to general information about personalized medicine and biomarkers and articles which do not provide further information on each broad of biomarker-guided designs were excluded. The 107 papers for biomarker-guided adaptive trial designs were reviewed in our published paper Antoniou et al. (2016) [35].

In our review, we identified five main biomarker-guided non-adaptive trial designs namely: (i) single-arm designs; (ii) enrichment designs; (iii) randomize-all designs; (iv) biomarker-strategy designs and (v) other designs. Within each main design several subtypes and extensions were also identified. Graphical representations of the main designs and subtypes are given in Figures 3.2-3.16. Graphical representations of the extensions are given in Figures B.1-B.4 included in Appendix B. The characteristics and methodology of the main design types and subtypes are discussed below and are summarized in Table 3.1, whilst information on the extensions are discussed in Appendix B. Furthermore, sample size formulae for each biomarker-guided design are provided in Table 3.2.



**Table 3.1.** Types of Biomarker guided non-adaptive designs proposed within the last ten years.

Types of Biomarker-Guided Non-Adaptive Trial Designs	Utility	Advantages	Limitations
<p><b>Single arm designs</b> (7 papers) [30, 36-41] (see Figure 3.2)</p> <p><b>Also called:</b> Nonrandomized clinical trial design, Uncontrolled Cohort Pharmacogenetic Study design</p> <p><b>Examples of actual trials:</b> None identified <sup>a</sup></p>	Useful for initial identification and/or validation of a biomarker.	<p>(A1) Considered as a simple statistical design as there is no need for randomization of patients.</p> <p>(A2) Simple logistics.</p> <p>(A3) Not complex statistical design</p> <p>(A4) In some cases, these designs may be viewed as ethical as all patients are given the opportunity to experience the experimental treatment. However, they may be viewed as unethical if the novel treatment does not benefit a subgroup of patients or causes adverse events.</p>	(L1) There is no distinction between prognostic and predictive biomarker as patients are not randomized to experimental and control treatment arms.

<p><b>Enrichment designs</b> (71 papers) [1, 4, 7-9, 11, 13, 15, 16, 18, 19, 21, 23, 25-33, 36, 42-86] (see <b>Figure 3.3</b>)</p> <p><b>Also called:</b> Targeted design, Selection design, Efficient Targeted design, Biomarker-Enrichment design, Marker-enrichment design, Gene enrichment design, Enriched design, Clinically enriched Phase III study design, Clinically Enriched Trial design, Biomarker-Enriched design, Biomarker Enriched design, Biomarker Selected trial design, Screening enrichment design, Randomized Controlled Trial (RCT) of test positive design, Population enrichment design</p>	<p>Useful when we aim to test the treatment effect only in biomarker-positive subset for which there is prior evidence that the novel treatment is beneficial, but the candidate biomarker requires prospective validation.</p> <p>Useful when it is not ethical to assign biomarker-negative patients to the novel treatment for which there is prior evidence that it will not be beneficial for this subpopulation, or that it will harm them.</p> <p>Recommended when both the cut-off point for determination of biomarker-status of patients and the analytical validity of a biomarker are well established.</p>	<p>(A5) Evaluates the effect of the experimental treatment in the biomarker-positive subgroup in a simple and efficient way.</p> <p>(A6) Provides clear information about whether the novel treatment is effective for the biomarker-positive subgroup, thus these designs can identify the best treatment for these patients and confirm the usefulness of the biomarker.</p> <p>(A7) Reduced sample size as the assessment of treatment effect is restricted only to biomarker-positive subgroup. Therefore, if the selected biomarker is “biologically correct” and reliably measured, the used enrichment strategy could result in a large saving of randomized patients.</p> <p>(A8) Enables rapid accumulation of efficacy data.</p>	<p>(L2) Do not assess whether the experimental treatment benefits the biomarker-negative patients, thus we cannot obtain information about this subgroup. Also unable to demonstrate whether the targeted treatment is beneficial in the entire study population.</p> <p>(L3) Do not inform us directly about whether the biomarker is itself predictive because the relative treatment efficacy may be the same in the unevaluated biomarker-negative patients. Since these designs only enrol a subgroup of patients, they do not allow for full validation of the marker’s predictive ability. For full validation, a trial would need to randomize all patients in order to test</p>
---	---	--	--

<p><b>Examples of actual trials:</b></p> <p>CRYSTAL [49], BRIM 3 [49-51], EURTAC [49], CLEOPATRA [49], PROFILE 1007 [49, 50], LUX-Lung [49], NSABP B-31 and NCCTG N9831 [4, 15, 16, 18, 19, 28-31, 36, 44, 46, 52-60], CALGB-10603 [61], CATNON [62], CODEL [62], Evaluation of epidermal growth factor receptor variant III (EGFRvIII) peptide vaccination [62], N0923 [7, 21], Flex study [64], TOGA trial [47], IPASS [33, 43], N0147 [29], PetaCC-8 [29, 47], C80405 [29], ECOG E5202 [29]</p>	<p>(A9) Allow us to avoid potential dilution of the results due to the absence of biomarker-negative patients. For example, if the design had included the biomarker-negative population and the biomarker positivity rate was low as compared to the biomarker negative rate, then the estimation of the overall treatment effectiveness could be diluted as it would be driven by the biomarker-negative subset.</p> <p>(A10) Can be attractive in terms of speed and cost, meaning that patients are provided with tailored treatment sooner.</p>	<p>for a treatment–biomarker interaction.</p> <p>(L4) Researchers should carefully decide whether or not to follow this strategy as it may be of limited value due to the exclusion of biomarker-negative patients. It may be that the entire population could benefit from the experimental treatment equally irrespective of biomarker status, in which case enrolling only the biomarker-positive patients will result in slow trial accrual, increase of expenses and unnecessary limitation of the size of the indicated patient population.</p> <p>(L5) Concern over an ethical problem as we cannot include individuals in a clinical trial if it is believed that the treatment is not effective for them, as</p>
--	--	---

raised by the US Food and Drug Administration (FDA) [50]. It was based on the facts that the experimental treatment can only be approved for a particular biomarker-defined subpopulation (i.e. biomarker-positive patients) if a companion diagnostic test is also approved, and how the test can be approved if the Phase III trial does not show that the novel treatment does not benefit the biomarker-negative patients.

(L6) The accuracy of diagnostic devices used to identify the biomarkers, e.g., biomarker assays, is not always correct [45]. This can result in incorrect selection of biomarker-positive patients and therefore these patients will erroneously be enrolled in a trial

			yielding biased treatment effect estimates. For example, even when the experimental treatment works well for a specific subgroup, if the biomarker assay is not able to identify this subgroup robustly then a promising treatment may be abandoned.
<p><b>Marker Stratified designs</b> (45 papers) [4, 10, 12, 13, 15-19, 21, 25-27, 30, 31, 33, 44-46, 49-51, 53, 58, 61, 62, 66, 68, 71-74, 79-81, 84-93] (see <b>Figure 3.4</b>)</p> <p><b>Also called:</b> Marker-stratified design, Biomarker-stratified design, Stratified-Randomized design, Stratification design, Stratified design, Stratified Analysis design, Marker by treatment – interaction design,</p>	<p>Useful when there is evidence that the novel treatment is more effective in the positive biomarker-defined subgroup than in the negative biomarker-defined subgroup but there is insufficient compelling data indicating that the experimental treatment does not benefit the biomarker-negative patients.</p>	<p>(A11) Ability to assess the treatment effect not only in the entire population but also in each biomarker-defined subgroup. Thus, this design can find the optimal treatment in the entire population and in each biomarker-defined subgroup.</p> <p>(A12) An ethical design even in situations where the biomarker is not useful as no treatment decisions are made based on biomarker status; all decisions are made randomly. Consequently, if the biomarker’s</p>	<p>(L7) In situations where there are several biomarkers and treatments this design may not be feasible as it involves randomization of patients between all possible treatment options and may require a large sample size.</p> <p>(L8) May not be feasible when the prevalence of the biomarker is low.</p>

<p>Marker-by-treatment interaction design, Treatment by marker interaction design, Treatment-by-marker interaction design, Marker <math>\times</math> treatment interaction design, Treatment-marker interaction design, Biomarker-by-treatment interaction design, Non-targeted RCT (stratified by marker) design, Genomic Signature stratified designs, Signature-Stratified design, Randomization or analysis stratified by biomarker status design, marker-interaction design.</p> <p><b>Examples of actual trials:</b></p> <p>MARVEL (N023) [4, 16, 30, 31, 33, 44, 61, 89], GALGB-30506 [15, 61], RTOG0825 [45], EORTC 10994 p53 [12, 66], IBCSG trial IX [18], MINDACT [18]</p>	<p>value is in doubt, this design may be preferred.</p>	<p>(L9) Might be expensive to test the entire population for its biomarker status.</p> <p>(L10) Measuring the biomarker upfront may be logistically difficult.</p> <p>(L11) There is no guarantee of balanced groups for analysis.</p>
--	---	--

<p><b><u>Sequential Subgroup-Specific design</u></b> (11 papers) [13, 14, 19, 22, 53, 57, 58, 60, 69, 91, 94] (see <b>Figure 3.5</b>)</p> <p><b>Also called:</b> sequential design, Fixed-sequence 2 design, hierarchical fixed sequence testing procedure</p> <p><b>Examples of actual trials:</b> PRIME [49], MARVEL [49]</p>	<p>Recommended when prior evidence indicates that the biomarker-positive subpopulation benefits more from the novel treatment as compared to the biomarker-negative subpopulation.</p>	<p>(A13) Allows for the estimation of treatment effect in biomarker-positive and biomarker-negative subgroups.</p> <p>(A14) Preserves the overall type I error rates and allows for a smaller sample size than the parallel version mentioned below.</p> <p>(A15) Considered as the best direct evidence for clinical decision making as it tests the treatment effectiveness in both the biomarker-positive and biomarker-negative subset in a sequential way.</p> <p>(A16) Do not require larger sample size than the overall/biomarker-positive designs when the prevalence of the biomarker-positive patients is small.</p>	<p>(L12) Has less power when there is homogeneity of treatment across the different biomarker defined subgroups as compared to the overall/biomarker-positive designs.</p> <p>(L13) Need a much larger sample size than the overall/biomarker positive designs if we assume that the treatment effect is relatively homogeneous across the biomarker-defined subsets.</p>
<p><b><u>Parallel Subgroup-Specific design</u></b> (3 papers) [14, 49, 69] (see <b>Figure 3.6</b>)</p>	<p>Appropriate when the aim of the study is to give treatment recommendations for each</p>	<p>(A17) Same as (A13), (A16)</p>	<p>(L14) Same as (L12)</p> <p>(L15) Allocates the overall level <math>\alpha</math> between the two biomarker-defined</p>

<p><b>Also called:</b> Phase III Biomarker-Stratified design</p> <p><b>Examples of actual trials:</b> None identified <sup>a</sup></p>	<p>biomarker-defined subgroup separately at the same time.</p>	<p>subgroup tests which means that it will be more difficult to achieve statistical significance in the biomarker-positive subgroup.</p>
<p><b><u>Biomarker-positive and overall strategies with parallel assessment</u></b> (8 papers) [1, 14, 36, 47, 49, 69, 95, 96] (<b>see Figure 3.7</b>)</p> <p><b>Also called:</b> Overall/biomarker-positive design with parallel assessment, prospective subset design, hybrid design</p> <p><b>Examples of actual trials:</b> S0819 [14, 49], SATURN [14, 36, 47, 49, 95, 96], MONET1 [14, 49], ARCHER [14, 49], ZODIAC [49], MERiDiAN [49]</p>	<p>Recommended when the aim of the study is to assess the treatment effect in both the entire population and in the biomarker-positive subset but not in the biomarker-negative population.</p>	<p>(A18) Can control the overall type I error <math>\alpha</math>.</p> <p>(A19) Can require smaller sample size as compared to the subgroup-specific designs, especially when we assume that the novel treatment equally benefits both biomarker-defined subgroups.</p> <p>(L16) Can be overly conservative as in the SATURN trial because of the correlation between the test of treatment effect in the overall study population and in the biomarker subgroups.</p> <p>(L17) Cannot control the probability of rejecting the null hypothesis of no treatment effect in the biomarker-negative subset when the treatment benefit is restricted to biomarker-positive patients. Consequently, there is a high risk of inappropriately recommending the novel treatment for biomarker-negative patients due</p>



			to the large treatment effect in biomarker-positive subset.
<p><b><u>Biomarker-positive and overall strategies with sequential assessment</u></b> (11 papers) [13, 14, 30, 44, 49, 69, 80, 84, 85, 88, 94] (see Figure 3.8)</p> <p><b>Also called:</b> Overall/biomarker-positive design with sequential assessment, sequential design, Fixed-sequence 2 design, hierarchical fixed sequence testing procedure</p> <p><b>Examples of actual trials:</b> Trial of letrozole plus lapatinib versus letrozole plus placebo in breast cancer, with the biomarker defined by human epidermal</p>	<p>Might be useful in cases where the experimental treatment is expected to be effective in the overall population.</p>	<p>(A20) Same as (A18), (A19)</p>	<p>(L18) Can be problematic for determining whether the treatment is beneficial in the biomarker-negative subgroup.</p> <p>(L19) Same as (L17)</p>

<p>growth factor receptor 2 (HER2)</p> <p>[14], N0147 [30, 49]</p>			
<p><b><u>Biomarker-positive and overall strategies with fall-back analysis</u></b> (15 papers) [10, 30, 36, 44, 47, 49, 53, 57, 60, 69, 84, 88, 94, 96, 97] (see <b>Figure 3.9</b>)</p> <p><b>Also called:</b> Biomarker-stratified design with fall-back analysis, fall-back design, prospective subset design, sequential design, other analysis plan design, Fallback design</p> <p><b>Examples of actual trials:</b> None identified <sup>a</sup></p>	<p>Recommended when there is insufficient confidence in the predictive value of the biomarker and the novel treatment is assumed to probably benefit all patients.</p>	<p>(A21) Can assess the treatment effect in the biomarker-positive patients, if no benefit is detected in the overall population.</p> <p>(A22) Same as (A18), (A19)</p>	<p>(L20) Same as (L17), (L18)</p>
<p><b><u>Marker Sequential test design</u></b> (4 papers) [14, 49, 69, 94] (see <b>Figure 3.10</b>)</p>	<p>Recommended when biomarkers with strong credentials are available and we have</p>	<p>(A23) Can provide clear evidence of treatment benefit in the biomarker-positive subgroup and in the biomarker-negative subgroup.</p>	<p>(L21) In situations where biomarker status is not available for some of the patients included in the study, this</p>

<p><b>Also called:</b> MaST design, hybrid design</p> <p><b>Examples of actual trials:</b> ECOG E1910 [14, 49]</p>	<p>convincing evidence that the novel treatment is more effective in biomarker-positive than in biomarker-negative patients.</p> <p>Appropriate when we can assume that the treatment will not be beneficial in the biomarker-negative subpopulation unless it is effective for the biomarker-positive subpopulation.</p>	<p>(A24) Enables sequential testing of the treatment effect in the entire study population and in the biomarker-defined subgroups to restrict testing of the treatment effect in the entire population when there is no significant result in the biomarker-positive subset, while controlling the appropriate type I error rates.</p> <p>(A25) Results in higher power as compared to the sequential subgroup-specific design in cases where the treatment effect is homogeneous across the biomarker-defined subgroups.</p> <p>(A26) Preserves the power in situations where the treatment effect is restricted only to the biomarker-positive patients and at the same time it controls the relevant type I error rates.</p> <p>(A27) Control the type I error rate for the biomarker-negative subgroup over all possible prevalence values.</p>	<p>design can either exclude these patients or include them in the global test, however, further statistical adjustments might be required in that case.</p> <p>(L22) Does not decrease the sample size of the study as it was developed in order to increase the power compared to the sequential subgroup-specific design in situations where the novel treatment benefits equally both biomarker-negative and biomarker-positive patients.</p>
--	---	---	---

		(A28) The probability of erroneously concluding that the novel treatment is beneficial for the entire population when the global effect is driven by the biomarker-positive patients is minimized since the design only tests the treatment effect in the entire population when no significant effect is detected in the biomarker-positive subgroup.	
<p><b>Hybrid designs</b> (14 papers) [1, 13, 15, 29-31, 36, 46, 48, 55, 66, 84, 88, 98] (see Figure 3.11)</p> <p><b>Also called:</b> Mixture design, Combination of trial designs, hybrid biomarker design</p> <p><b>Examples of actual trials:</b>          TAILORx [15, 48, 55, 58, 63, 66],          EORTC MINDACT [15, 48, 55, 66],          ECOG 5202 study [30, 46]</p>	<p>Can be used when there is prior evidence indicating that only a particular treatment is beneficial to a biomarker-defined subgroup which makes it unethical to randomize patients with that specific biomarker status to other treatment options.</p>	<p>(A29) The feasibility of a prognostic biomarker can be tested.</p> <p>(A30) Allows for better risk assessment and improved individualized treatment since it assigns patients to treatments based on risk assessment scores instead of their biomarker status (biomarker-positive and biomarker-negative patients).</p>	<p>None found.</p>

<p><b><u>Biomarker-strategy designs with biomarker assessment in the control arm</u></b> (21 papers) [15, 25, 26, 32, 33, 36, 45, 61, 62, 64, 79, 82, 85, 86, 92, 93, 99-103] (see Figure 3.12)</p> <p><b>Also called:</b> Marker strategy design, Biomarker-strategy design, Strategy design, Marker-based strategy design, Marker-based design, Random disclosure design, Customized strategy design, Parallel controlled pharmacogenetic study design, Marker-based strategy design I, Biomarker-guided design, Biomarker-based assignment of specific drug therapy design, Marker-based strategy I design, Biomarker-strategy design with a</p>	<p>Useful when we want to test the hypothesis that the treatment effect based on the personalized approach is superior to that of the standard of care.</p>	<p>(A31) Biomarker can be validated without including all possible biomarker–treatment combinations [26] as in the non-biomarker-based arm all patients receive only the control treatment.</p> <p>(A32) Have the option of testing the biomarker status of patients in the non-biomarker-strategy arm which can aid secondary analyses [26].</p> <p>(A33) Able to inform us whether the biomarker is prognostic.</p> <p>(A34) Can be expanded to investigate several biomarkers and treatments [103]. Additionally, these designs can be attractive when evaluating multiple biomarkers or the predictive value of molecular profiling between several treatment options is to be assessed [45].</p>	<p>(L23) Unable to inform us whether the biomarker is predictive as these designs are able to answer the question about whether the biomarker-based strategy is more effective than standard treatment, irrespective of the biomarker status of the study population.</p> <p>(L24) The evaluation of the true biomarker by treatment effect is not possible as the biomarker-positive patients receive only the experimental treatment and not the alternative treatment (control treatment). Consequently, this design cannot detect the case in which the control treatment might be more beneficial for the entire population.</p> <p>(L25) In case that the number of biomarker-positive patients is very</p>
---	---	---	---

<p>standard control, Marker strategy design for prognostic biomarkers</p> <p><b>Examples of actual trials:</b> GILT docetaxel [15], Randomized phase III trial conducted in Spain, dedicated to patients with advanced Non-Small Cell Lung Cancer (NSCLC) candidates for first-line chemotherapy [32, 64, 100], Study the effect of Magnetic Resonance Imaging (MRI) in patients with low back pain on patient outcome and to evaluate Doppler US of the umbilical artery in the management of women with intrauterine growth retardation (IUGR), Randomized controlled trial in recurrent platinum-resistant ovarian carcinoma [101]</p>	<p>(A35) Might be used more frequently in the future due to the wide variety of molecular biomarkers, complexity of gene expression arrays, and several treatments directed at similar targets [103].</p>	<p>small, then the treatment received will be similar in biomarker-strategy arm and non-biomarker strategy arm. Consequently, the trial might give little information regarding the efficacy of the experimental treatment or it might not be able to detect it. As a result, this type of design should be used when there is an adequate number of biomarker-positive and biomarker-negative patients.</p> <p>(L26) Unable to compare directly experimental treatment to control treatment as the aim is to compare not the treatments but the biomarker-strategies.</p> <p>(L27) Less efficient designs than biomarker-stratified designs [4, 73] and a poor substitute for clinical trials which aim to compare the</p>
---	---	---

experimental treatment to control treatment, since it is possible for some patients in both the biomarker-based strategy arm and non-biomarker-based strategy arm to be assigned to the same treatment (due to the existence of biomarker-negative patients in both strategy arms the treatment effect can be diluted) [51]. Consequently, as a large overlap of patients receiving the same treatment might have occurred, the comparison of the two biomarker-strategy arms results in a hazard ratio which is forced towards unity, i.e. no treatment effect exists as the effect of experimental versus control treatment is diluted by the biomarker-based treatment selection. For this reason, a large sample size is needed to detect at least a small overall difference in

			<p>outcomes between the two biomarker-strategy arms.</p> <p>(L28) Should be used only if you want to evaluate a complex biomarker-guided strategy with a variety of treatment options or biomarker categories [73].</p>
<p><b><u>Biomarker-strategy design without biomarker assessment in the control arm</u></b> (14 papers) [9, 13, 17, 18, 20, 25, 36, 38, 61, 74, 101, 104-106] (see Figure 3.13)</p> <p><b>Also called:</b> Biomarker-strategy design with standard control, Direct-predictive biomarker-based, RCT of testing, Test-treatment, Parallel controlled pharmacogenetic diagnostic study, Marker strategy, Marker-</p>	<p>In situations where it is not feasible or unethical to test the biomarker in the entire population.</p>	<p>(A36) Galanis et al., 2011 [45] stated that these designs can be attractive when evaluating multiple biomarkers or the predictive value of molecular profiling between several treatment options is to be assessed. Also, Freidlin and Korn, 2010 [73] claimed that these biomarker-strategy designs should be used only if researchers want to evaluate a complex biomarker-guided strategy with a variety of treatment options or biomarker categories.</p> <p>(A37) Same as (A31), (A32), (A33)</p>	<p>(L29) Criticized for their potential cost increase due to the fact that patients without predicted responsive biomarker are double enrolled in the trial (biomarker-negative patients receive control treatment in both strategy arms).</p> <p>(L30) Biomarker-positive and biomarker-negative subpopulations might be more imbalanced as compared with the first type of biomarker-strategy design due to the</p>



<p>based with no randomization in the non-marker-based arm, Classical, Marker-based strategy, Marker strategy design for prognostic biomarkers</p> <p><b>Examples of actual trials:</b> A study, which evaluated the use of immediate computed tomography in patients with acute mild head injury [101, 104].</p>			<p>fact that the randomization to different treatment strategies is performed before the evaluation of the biomarker status (balancing the randomization is useful to ensure that all randomized patients have tissue available). This can happen especially when the number of patients is very small.</p> <p>(L31) Same as (L23), (L24), (L25), (L26), (L27)</p>
<p><b><u>Biomarker-strategy design with treatment randomization in the control arm</u></b> (17 papers) [15, 17, 26, 27, 32, 36, 45, 62, 64, 66, 74, 86, 92, 93, 106-108] (see Figure 3.14)</p> <p><b>Also called:</b> Biomarker-strategy design with a randomized control, Modified marker-based strategy</p>	<p>In cases where we want to know whether the biomarker is not only prognostic but also predictive, these designs are preferable as compared to the two previously mentioned biomarker-strategy designs.</p>	<p>(A38) These designs have the ability to inform researchers about the potential superiority of the control treatment in the whole population or among a particular biomarker-defined subpopulation.</p> <p>(A39) Able to inform us whether the biomarker is prognostic or predictive.</p>	<p>(L32) Generally require a larger sample size as compared to the marker-stratified designs.</p> <p>(L33) Same as (L27)</p>

<p>design (for predictive biomarkers), Biomarker-strategy design with randomized control, Marker- based design with randomization in the non-marker-based arm, Marker-based strategy design II, Marker-strategy design, Augmented strategy design, Trial design allowing the evaluation of both the treatment and the marker effect</p> <p><b>Examples of actual trials:</b> None identified <sup>a</sup></p>		<p>(A40) Allow clarification of whether the results which indicate efficacy of the biomarker-directed approach to treatment are caused due to a true effect of the biomarker status or to an improved treatment irrespective of the biomarker status.</p> <p>(A41) Same as (A36)</p>	
<p><b>Reverse marker-based strategy</b> (4 papers) [86, 92, 93, 109] (see <b>Figure 3.15</b>)</p> <p><b>Also called:</b> None found</p>	<p>Enables testing the interaction hypothesis of treatment and biomarker in a more efficient way as compared to the first (i.e. Biomarker-strategy design with biomarker assessment in the control arm) and third</p>	<p>(A42) Can estimate directly the marker- strategy response rate.</p> <p>(A43) Allows the estimation of the effect size of the experimental treatment compared to the control treatment for each biomarker- defined subset separately.</p>	<p>(L34) It has been claimed by Baker, 2014 [93] that other designs than the reverse marker-based strategy are more appropriate in order to investigate questions which include both treatment effect of biomarker- defined subgroups and the biomarker</p>

<p><b>Examples of actual trials:</b> None identified <sup>a</sup></p>	<p>biomarker-strategy subtype design (i.e. Biomarker-strategy design with randomization in the control arm and the marker stratified design)</p>	<p>(A44) There is no chance that the same treatment will be tailored to biomarker-positive patients who are randomized either to the biomarker-based strategy arm or the reverse marker strategy. Also, there is no possibility of the same treatment assignment to biomarker-negative patients who are randomly assigned to the two biomarker-based strategy arms.</p> <p>(A45) It has been demonstrated by Eng, 2014 [92] that this new type of design is more than four times more efficient for testing the interaction between treatment and biomarker compared to Biomarker-strategy design with biomarker assessment in the control arm, Biomarker-strategy design with randomization in the control arm and the marker stratified design.</p>	<p>strategy treatment effect. These designs should allow the estimation of treatment effects within biomarker-defined subgroups as well as the estimation of the global treatment effect.</p>
---	--	---	---

<b><u>A specific randomized phase II trial design that can be used to guide decision making for further development of an experimental therapy.</u></b> (1 paper) [71] (see Figure 3.16)	Recommended when we want to conduct a Phase II randomized trial which allows decisions to be made about which type of Phase III biomarker-guided trial should be used.	(A46) Works well in providing recommendations for phase III trial design.	None found
--	--	---	------------

**Table 3.2.** Sample size formulae for biomarker-guided clinical trial designs.

<b>Types of Biomarker-Guided Non-Adaptive Trial Designs</b>		<b>Sample Size Formula</b>	<b>Definition</b>
<b><u>Single arm designs</u></b>		Standard sample size formula can be used, more information can be found in the ‘methodology’ part of the ‘Single arm designs’ section in the main text.	
<b><u>Enrichment designs</u></b> [55, 61, 65, 110-112]		Online tool for sample size calculation when using either binary or time-to-event endpoints is available on the following website: <a href="http://brb.nci.nih.gov/brb/samplesize/td.html">http://brb.nci.nih.gov/brb/samplesize/td.html</a> [113].	

---


$$E(D_{i, enrichment}) = \frac{nT\lambda_i}{2(\lambda_i + \varphi_i)} \left\{ 1 - \frac{e^{-(\lambda_i + \varphi_i)\tau}}{(\lambda_i + \varphi_i)T} [1 - e^{-(\lambda_i + \varphi_i)T}] \right\}$$

$E(D_{i, enrichment})$  refers to the expected number of events per treatment arm (time-to-event outcome),  $i$  corresponds to either the experimental or the control treatment group, 1: 1 ratio between the two treatment arms (experimental: control) is assumed,  $\lambda$  corresponds to the event hazard rate,  $\varphi$  is the loss to follow-up rate,  $T$  denotes the accrual time, patients enter the trial according to a Poisson process with rate  $n$  per year over the accrual period of  $T$  years,  $\tau$  corresponds to the follow-up period.

---


$$D_{enrichment} = 4 \left[ \frac{(z_{\alpha/2} + z_{\beta})^2}{\log \theta_1} \right]^2$$


---

$D_{enrichment}$  refers to the required total number of events (time-to-event outcome), 1: 1 ratio between the two treatment arms (experimental:control) is assumed,

---

$z_{\alpha/2}, z_{\beta}$  denote the upper  $\alpha/2$ - and upper  $\beta$ -points respectively of a standard normal distribution,  $\alpha$  and  $\beta$  denote the assumed type I error and type II error respectively,  $\theta_1$  denotes the assumed hazard ratio between the two treatment groups (control vs experimental) in the biomarker-positive subset.

---

$$N_{enrichment/arm} = 2\bar{p}_Q(1 - \bar{p}_Q) \left[ \frac{(z_{\alpha/2} + z_{\beta})}{(p_A^Q - p_B)} \right]^2$$

---

$N_{enrichment/arm}$  refers to the required number of patients per treatment arm (binary outcome), 1:1 ratio between the two treatment arms (experimental: control) is assumed,  $p_A^Q$  and  $p_B$  are the response probabilities in the experimental and control groups respectively,  $\bar{p}_Q = (p_A^Q + p_B)/2$ .  $p_A^Q = p_B + \delta_+$ , where  $\delta_+$  denotes the improvement in response

---

	probability for biomarker-positive patients.
$N_{enrichment/arm} = \frac{2\sigma^2(z_{a/2} + z_\beta)^2}{(\mu_{A+} - \mu_{B+})^2}$	<p><math>N_{enrichment/arm}</math> refers to the required total number of patients per treatment arm (continuous response endpoints), 1: 1 ratio between the two treatment arms (experimental: control) is assumed, <math>\sigma^2</math> denotes the anticipated common variance, <math>\mu_{A+}</math> and <math>\mu_{B+}</math> the mean responses for biomarker-positive patients in the experimental and control treatment arm respectively.</p>
$N_{enrichment/arm} = 2\sigma^2(z_{a/2} + z_\beta)^2 \{\lambda_1[(1 - \omega)\zeta + \omega]\}^{-2}$	<p><math>N_{enrichment/arm}</math> refers to the required total number of patients per treatment arm (continuous response endpoints when accounting for error in the assaying of the study population), 1: 1 ratio between the two treatment arms</p>

		(experimental:control) is assumed, $\omega$ measures the accuracy of the assay and corresponds to the PPV (positive predictive value of the assay, i.e. the proportion of patients who are assigned biomarker positive status according to the assay who are truly biomarker positive), $\lambda_1$ is the treatment effect in the biomarker-positive patients and $\zeta = \lambda_0/\lambda_1$ (where $\lambda_0$ is the treatment effect in the biomarker-negative patients).
<b><u>Marker Stratified designs</u></b>	Online tool for sample size calculation when using either binary or time-to-event endpoints is available on the following website: <a href="http://brb.nci.nih.gov/brb/samplesize/sdpap.html">http://brb.nci.nih.gov/brb/samplesize/sdpap.html</a> [115].	
[31, 53, 60, 92, 111, 112, 114]		
	$D_{stratified} = 4 \frac{(z_{a_1} + z_{\beta})^2}{[\log(\theta_1)]^2} + 4 \frac{(z_{a_2} + z_{\beta})^2}{[\log(\theta_2)]^2}$	$D_{stratified}$ refers to the required total number of events for the achievement of sufficient power in each biomarker-defined subgroup



	separately (time-to-event endpoint), 1: 1 ratio between the two treatment arms (experimental: control) is assumed, $\theta_2$ corresponds to the hazard ratio of biomarker-negative subgroup, $a_1 = a_2 = a/2$ .
$D_{stratified} = \frac{4(z_{a/2} + z_{\beta})^2}{[k\log(\theta_1) + (1 - k)\log(\theta_2)]^2}$	$D_{stratified}$ refers to the required total number of events for the achievement of sufficient power in the overall population (time-to-event endpoint), $k$ is the proportion biomarker-positive patients, 1: 1 ratio between the two treatment arms (experimental: control) is assumed.
$N_{stratified} = \frac{4(z_{a/2} + z_{\beta})^2}{\{[kPr_{(+)}(event)\log(\theta_1) + (1 - k)Pr_{(-)}(event)\log(\theta_2)]/\sqrt{kPr_{(+)}(event) + (1 - k)Pr_{(-)}(event)}\}^2}$	$N_{stratified}$ refers to the required total number of patients for the achievement of sufficient power in the overall population (time-to-event endpoint), 1: 1 ratio between

	<p>the two treatment arms</p> <p>(experimental: control) is assumed,</p> <p><math>Pr_{(+)}(event)</math>, <math>Pr_{(-)}(event)</math> are the probabilities of an event in biomarker-positive subset and biomarker-negative subset respectively.</p>
$\frac{D_{stratified}}{D_{enrichment}} = \frac{[\log(\theta_1)]^2}{[k\log(\theta_1) + (1-k)\log(\theta_2)]^2} = \frac{1}{\left[k + (1-k)\frac{\log(\theta_2)}{\log(\theta_1)}\right]^2}$	<p><math>\frac{D_{stratified}}{D_{enrichment}}</math> refers to the ratio of the required number of events between marker stratified and enrichment design (time-to-event endpoint).</p>
$\frac{N_{stratified}}{N_{enrichment}} \approx \frac{1}{\left[k + (1-k)\frac{\delta_-}{\delta_+}\right]^2}$	<p><math>\frac{N_{stratified}}{N_{enrichment}}</math> refers to the ratio of the required number of patients between marker stratified and enrichment design (binary outcome), <math>\delta_-</math>, <math>\delta_+</math>, correspond to the treatment effectiveness in biomarker-negative and</p>

	<p>biomarker-positive subgroup respectively.</p>
$N_{stratified} = 2(z_a + z_{1-\beta})^2 \left\{ \frac{r_{A+}(1 - r_{A+}) + r_{B+}(1 - r_{B+})}{(\beta_A + \beta_I)^2} + \frac{r_{A-}(1 - r_{A-}) + r_{B-}(1 - r_{B-})}{(\beta_A)^2} \right\}$	<p><math>N_{stratified}</math> refers to the required total number of patients (binary outcome), <math>\beta_0</math> denotes a baseline effect, <math>\beta_A</math> denotes the added effect of the experimental treatment, <math>\beta_+</math> denotes the biomarker-positive effect and <math>\beta_I</math> denotes the nonadditive effect, <math>a</math> corresponds to the target level, <math>1 - \beta</math> corresponds to the power, <math>r_{A+}, r_{B+}</math> are the assumed response rates of biomarker-positive patients receiving the experimental and the control treatment respectively, <math>r_{A-}, r_{B-}</math> are the assumed response rates of biomarker-negative patients receiving the experimental</p>

		and the control treatment respectively.
<b><u>Sequential Subgroup-Specific design</u></b> [57]	$N_{\text{sequential subgroup-specific}}^+ = N_{\text{enrichment}}$	$N_{\text{sequential subgroup-specific}}^+$ refers to the required number of biomarker-positive patients (binary outcome), $N_{\text{enrichment}}$ is the required number of biomarker-positive patients (binary outcome) in the enrichment design.
	$N_{\text{sequential subgroup-specific}} = \frac{N_{\text{enrichment}}}{k}$	$N_{\text{sequential subgroup-specific}}$ refers to the required total number of patients (binary outcome), $N_{\text{enrichment}}$ is the required number of biomarker-positive patients (binary outcome) in the enrichment design.
	$N_{\text{sequential subgroup-specific}}^- = \frac{(1-k)N_{\text{enrichment}}}{k}$	$N_{\text{sequential subgroup-specific}}^-$ refers to the required number of biomarker-negative patients (binary outcome),

	<p><math>N_{enrichment}</math> is the required number of biomarker-positive patients (binary outcome) in the enrichment design.</p>
$D_{sequential\ subgroup-specific}^+ = D_{enrichment}$	<p><math>D_{sequential\ subgroup-specific}^+</math> refers to the required number of events for biomarker-positive patients (time-to-event outcome), <math>D_{enrichment}</math> is the required number of events for biomarker-positive patients (time-to-event outcome).</p>
$D_{sequential\ subgroup-specific}^- = D_{enrichment} \left( \frac{\lambda_-}{\lambda_+} \right) \left( \frac{1-k}{k} \right)$	<p><math>D_{sequential\ subgroup-specific}^-</math> refers to the required number of events for biomarker-negative patients (time-to-event outcome), <math>D_{enrichment}</math> is the required number of events for biomarker-positive patients (time-to-event outcome), <math>\lambda_-</math>, <math>\lambda_+</math>, are the event rates in biomarker-negative</p>

	and biomarker-positive control subgroups.
<b><u>Parallel Subgroup-Specific design</u></b>	<p>Same formula proposed for marker stratified designs could be considered to achieve sufficient power in each biomarker-defined subgroup simultaneously. However, in order to control the overall type I error rate of the design at the overall level of significance <math>\alpha</math> it is required to allocate this overall <math>\alpha</math> between the test for the biomarker-positive subgroup and the test for the biomarker-negative. Consequently, for biomarker-positive subgroup the reduced significance level <math>\alpha_1 = \alpha - \alpha_2</math> can be used whereas the reduced significance level <math>\alpha_2 = \alpha - \alpha_1</math> can be used for biomarker-negative subgroup.</p>
<b><u>Biomarker-positive and overall strategies with parallel assessment</u></b>	<p>If there is significant confidence that the biomarker is predictive, the sample size estimation is aimed at having a sufficient number of biomarker-positive individuals to enable the treatment effect in the biomarker positive subgroup to be detected. Standard formula for sample size calculation of biomarker-positive subgroup proposed for the enrichment designs could be considered by using the reduced significance level <math>\alpha_1 = \alpha - \alpha_2</math>. On the other hand, if there is no confidence in the predictive value of the biomarker, the sample size estimation is aimed at having a sufficient number of patients to detect a treatment effect in the overall study population; consequently, for the sample size calculation, the same formula proposed for marker stratified designs aiming to achieve sufficient power in the overall population could be applied by using the reduced significance level <math>\alpha_2 = \alpha - \alpha_1</math>.</p>

<b><u>Biomarker-positive and overall strategies with sequential assessment</u></b>	At the first stage, the standard formula for a traditional randomized trial which is the same with the formula proposed for enrichment designs can be applied for the biomarker-positive subgroup. At the second stage, the sample size formula proposed for marker stratified designs aiming to yield appropriate power for the entire population could be considered.
<b><u>Biomarker-positive and overall strategies with fall-back analysis</u></b>	At the first stage, the sample size formula proposed for marker stratified designs aiming to yield appropriate power for the entire population could be considered by using the reduced significance level $\alpha_1 = \alpha - \alpha_2$ . At the second stage, the formula proposed for enrichment designs could be applied for the biomarker-positive subgroup by using the reduced significance level $\alpha_2 = \alpha - \alpha_1$ .
<b><u>Marker Sequential test design (MaST)</u></b>	A standard sample size calculation (i.e. the same sample size calculation as for the enrichment designs) can be applied for the biomarker-positive subpopulation. However, in order to have sufficient number of biomarker-positive patients to detect treatment effectiveness in that particular biomarker-defined subset and consequently to reach the desired power, the sample size should be calculated by using the reduced significance level $\alpha_1 [0, \alpha]$ instead of the global significance level $\alpha$ which is used in the sample size formulae of the enrichment designs. The same formula could be considered for the sample size calculation of the biomarker-negative subgroup; however, the corresponding hazard ratio of that subgroup and the global significance level $\alpha$ should be used. For the sample size calculation of the entire population, the same

---

formula proposed for marker stratified designs aiming to achieve sufficient power in the overall population could be considered by using the reduced significance level  $\alpha_2 = \alpha - \alpha_1$ .

---

**Biomarker-strategy,  
design with biomarker  
assessment in the control  
arm** [26, 61, 92]

$$D_{strategy\,I} = 4 \left[ \frac{(z_{\alpha/2} + z_{\beta})}{k \log \theta_1} \right]^2$$

$D_{strategy\,I}$  refers to the required total number of events (time-to-event outcome), 1:1 ratio between the two treatment arms (experimental: control) is assumed.

$$N_{strategy\,I} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 (\tau_m^2 + \tau_n^2)}{(v_m - v_n)^2}$$

$N_{strategy\,I}$  refers to the required total sample size (continuous clinical endpoints), 1:1 ratio between the two treatment arms (experimental:control) is assumed,  $z_{1-\alpha/2}$ ,  $z_{1-\beta}$  denote the lower  $1 - \alpha/2$ - and lower  $1 - \beta$ -points respectively of a standard normal distribution,  $v_m$  and  $v_n$  denote the mean response from the biomarker-based strategy arm and the non-biomarker-based strategy arm respectively, and  $\tau_m^2$ ,  $\tau_n^2$

---



	denote the variance of response for the biomarker-based strategy arm and non-biomarker-based strategy arm respectively.
$N_{strategy\ 1/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_2(1 - g_2)]}{\Delta_2^2}$	<p><math>N_{strategy\ 1/arm}</math> refers to the required total number of patients per arm (binary outcome), <math>g_1</math> is the expected response rate in the biomarker-based strategy arm, <math>g_2</math> is the expected response rate in the non biomarker-based strategy arm, <math>\Delta_2 = g_1 - g_2</math>, <math>g_1, g_2</math> can be found by calculating the formulae <math>kr_{A+} + (1 - k)r_{B-}</math> and <math>r_B</math> respectively, <math>r_B</math> denotes the marginal effect of treatment B (control treatment).</p>
<p><b><u>Biomarker-strategy design without biomarker</u></b></p>	<p>Same formulae as for the ‘Biomarker-strategy design with biomarker assessment in the control arm’ can be considered.</p>

---

assessment in the control

arm

---

Biomarker-strategy design

with treatment

randomization in the

control arm [26, 31, 92]

$$D_{strategy III} = \frac{4(z_{a/2} + z_{\beta})^2}{\left\{ \log \left[ \frac{2km_{B+} + 2(1-k)m_{A-}}{k(m_{A+} + m_{B+}) + (1-k)(m_{A-} + m_{B-})} \right] \right\}^2}$$

$D_{strategy III}$  refers to the required total number of events (time-to-event outcome), 1:1 ratio between the two treatment arms (experimental: control) is assumed,  $m_{A+}, m_{A-}, m_{B+}, m_{B-}$ , denote the median survival for biomarker-positive and biomarker-negative patients receiving control and experimental treatments respectively.

$$N_{strategy III} = \frac{2(z_{1-a/2} + z_{1-\beta})^2 (\tau_m^2 + \tau_{nr}^2)}{(v_m - v_{nr})^2}$$

$N_{strategy III}$  refers to the required total sample size (continuous clinical endpoints), 1:1 ratio between the two treatment arms (experimental: control) is assumed,  $v_{nr}$  denotes the mean response from the non-biomarker-based

		strategy arm, $\tau_{nr}^2$ denotes the variance of response for the non-biomarker-based strategy arm respectively.
	$N_{strategy\ III/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_3(1 - g_3)]}{\Delta_3^2}$	$N_{strategy\ III/arm}$ refers to the required total number of patients per arm (binary outcome), $g_3$ is the expected response rate in the non biomarker-based strategy arm and $\Delta_3 = g_1 - g_3$ , the expected response rate $g_3$ can be found by calculating the formula $r_A/2 + r_B/2$ , $r_A$ denotes the marginal effect of treatment A (experimental treatment).
Reverse marker-based strategy [92]	$N_{strategy\ IV/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_4(1 - g_4)]}{\Delta_4^2}$	$N_{strategy\ IV/arm}$ refers to the required total number of patients per arm (binary outcome), $g_4$ is the expected response rate in the reverse biomarker-based strategy arm and $\Delta_4 = g_1 - g_4$ , the expected

---

response rate  $g_4$  can be found by calculating the formula  $kr_{B+} + (1 - k)r_{A-}$ ,  $r_{B+}$ ,  $r_{A-}$  are the assumed response rates of biomarker-positive patients receiving the control treatment and biomarker-negative patients receiving the experimental treatment.

---

**Randomized Phase II trial design with biomarkers**

[71]

Online tool for sample size calculation is available on the following website:  
<http://brb.nci.nih.gov/Data/FreidlinB/RP2BM> [115].

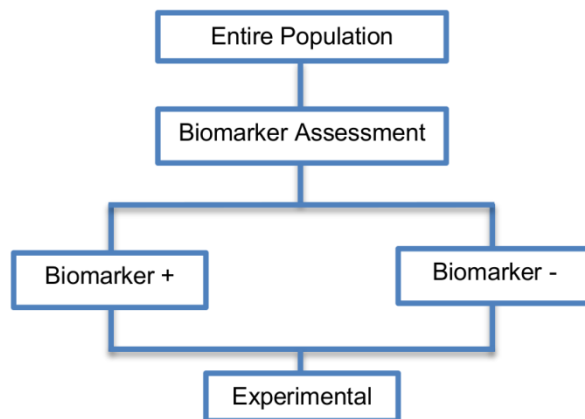
---

### 3.2.1. Single Arm Designs

---

Single arm designs were referred to in seven papers (7%). In the context of biomarkers, these designs (Phase II designs) include the whole study population to which the same experimental treatment is prescribed, without taking into consideration biomarker status.

**Design:** In this design all patients are prescribed the experimental treatment and there is no comparison with a control treatment. These trial designs aid in the identification of association between biomarker status and the efficacy or safety of the experimental treatment. An illustration of this approach is shown in Figure 3.2.



**Figure 3.2.** Single arm designs

**Utility:** These designs can be useful for the initial identification and/or validation of a biomarker and their aim is not to estimate the treatment effect in a definitive way but to identify whether the biomarker is sufficiently promising to proceed to a definitive Phase III biomarker-guided randomized controlled trial.

**Methodology:** In single arm designs first, we assess the biomarker status of patients and then as all patients will be treated the same way we could compare the outcome of the biomarker-positive subgroup with the outcome of biomarker-negative subgroup. According to Tajik et al., 2012 [116], in terms of the required sample size, a standard formula can be used, however one should take into consideration the multiple testing issue that arise due to the exploration of several

prognostic biomarkers (e.g., Bonferroni adjustment or normal exact method to protect against type I error  $\alpha$  for multiple tests are often considered [117]). Further information can be found in the paper of Zaslavsky and Scott, 2012 [117] who studied the sample size estimation in single arm clinical trials with multiple testing under frequentist and Bayesian framework.

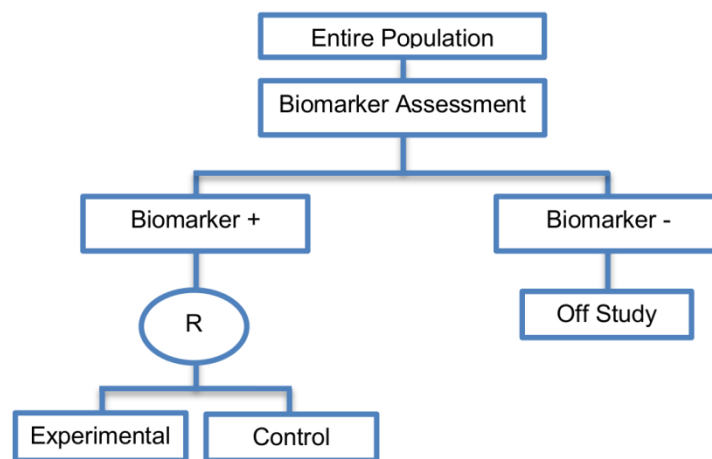
**Statistical considerations:** The single arm approach can be considered as a simple statistical design as there is no need for randomization. However, one limitation of this strategy is that there is no distinction between prognostic and predictive biomarkers i.e. as patients are not randomized to experimental and control treatment groups, it is not possible to determine whether an observed effect is attributable to the natural disease progression or to the treatment. Consequently, this study designs are unable to show the benefit of a biomarker with regard to the best choice of treatment.

### 3.2.2. Enrichment Designs

---

Enrichment designs are described in 71 papers (71%), either in Phase II or Phase III clinical trials, and involve randomizing only the biomarker-positive patients and comparing the experimental treatment versus the standard treatment only in this particular biomarker-defined subgroup.

**Design:** Figure 3.3 graphically represents the trial design. First, the entire population is screened in order to identify the biomarker status of each patient. Next, the random assignment of individuals to different treatment arms is restricted only to the biomarker-positive subgroup. More precisely, biomarker-negative patients are excluded from the study and consequently, the assessment of the effectiveness of the experimental treatment is limited to the biomarker-positive subgroup. Thus, other patients apart from the biomarker-positive subpopulation can receive only the standard treatment (i.e. control treatment), but they are not included in the investigation during the trial design. The biomarker in this design is referred to as either the 'selection' or 'enrichment' biomarker.



**Figure 3.3.** Enrichment designs. “R” refers to randomization of patients.

**Utility:** Enrichment designs are useful for clinical trials aiming to test the treatment effect in a specific biomarker-defined subpopulation where there is evidence to suggest that effectiveness is limited to those within that subgroup, but the candidate biomarker still requires prospective validation. This design is recommended when both the cut-off point for determination of biomarker status of patients and the analytical validity of the biomarker have been well established. A rapid turnaround time for assessing the biomarker status of a patient is also needed to avoid any delay in treatment initiation. This strategy is particularly useful where it is unethical to randomize the biomarker-negative population into different treatment arms, for example where there is prior evidence that the experimental treatment is not beneficial for biomarker-negative individuals, or is likely to cause them harm. However, when it remains unclear whether or not biomarker-negative individuals will benefit from the novel treatment, the enrichment design is not appropriate and alternative designs, which also assess effectiveness in the biomarker-negative individuals, should be considered (e.g., randomize-all designs).

**Methodology:** An online tool has been developed by Zhao and Simon [19, 28, 53, 57, 60] that allows sample size planning for the enrichment design both for binary and time-to-event (survival) outcomes, and is available at <http://brb.nci.nih.gov/brb/samplesize/td.html> [113]. For the purpose of estimating the sample size in the case of a survival outcome, data are simulated based on a marker stratified design (see next section for further information) in which both biomarker-

positive and biomarker-negative subgroups are investigated in the study and formulae for the enrichment design described in the paper of Rubinstein et al., 1981 [110] are used. Furthermore, an exponential distribution of survival for the experimental and control treatment groups within both the biomarker-positive and biomarker-negative subpopulations is assumed. More precisely, Rubinstein et al. provide the formula of the expected number of events per treatment group allowing to include exponential loss to follow-up given the following assumptions: (i) patients enter the trial according to a Poisson process and patient entry times will be independent and identically distributed uniformly over  $[0, T]$  where  $T$  denotes the accrual time. Consequently, given the total number of patients  $N$ , the times from entry to the end of the trial will be independent and identically distributed uniformly over  $[\tau, T + \tau]$ , where  $\tau$  denotes the follow-up time and  $T + \tau$  the total duration of the study and (ii) 1:1 randomization between experimental and control treatment group is considered. The expected number of events per treatment arm according to Rubinstein et al. is given by

$$E(D_{i, \text{enrichment}}) = \frac{nT\lambda_i}{2(\lambda_i + \phi_i)} \left\{ 1 - \frac{e^{-(\lambda_i + \phi_i)t}}{(\lambda_i + \phi_i)T} [1 - e^{-(\lambda_i + \phi_i)T}] \right\}, \quad (3.1)$$

where  $i$  corresponds to either the experimental or the control treatment group,  $\lambda$  corresponds to the event hazard rate,  $\phi$  is the loss to follow-up rate and patients enter the trial according to a Poisson process with rate  $n$  per year over the accrual period of  $T$  years. However, the required total number of events in the two treatment groups (experimental and control treatment group) is given by

$$D_{\text{enrichment}} = 4 \left[ \frac{(z_{a/2} + z_\beta)}{\log \theta_1} \right]^2, \quad (3.2)$$

where  $\theta_1$  denotes the assumed hazard ratio between the two treatment groups (control vs. experimental) in the biomarker-positive subset and the constants  $z_{a/2}, z_\beta$  denote the upper  $a/2$ - and upper  $\beta$ -points respectively of a standard normal distribution where  $a$  and  $\beta$  denote the assumed type I error and type II error respectively. Freidlin et al., 2010 [61] provided the aforementioned formula assuming



that all random assignments use 1:1 randomization. As in a traditional randomized controlled trial, if the randomization is not equal, i.e. the ratio of allocation to treatment and control is  $R:1$  rather than 1:1, the aforementioned formula for the required total number of events  $D_{enrichment}$  which assumes 1:1 randomization can be multiplied by  $(R + 1)^2/4R$  [118]. Consequently, the “4” in the formula of  $D_{enrichment}$  becomes  $(R + 1)^2/R$  and the corresponding formula for the total number of events becomes

$$D_{enrichment} = \frac{(R + 1)^2}{R} \left[ \frac{(z_{\alpha/2} + z_{\beta})}{\log \theta_1} \right]^2. \quad (3.3)$$

In a survival study, the calculation of the total sample size in terms of number of patients required in the two treatment groups (experimental and control treatment group) to be enrolled in order to yield the aforementioned total number of events depends on the probability of event over the duration of the study [119]. Consequently, the actual number of patients required in a survival study can be given by

$$N_{enrichment} = \frac{D_{enrichment}}{\Pr(event)}, \quad (3.4)$$

where  $\Pr(event)$  is the probability of observing an event in the two treatment groups in the study and  $D_{enrichment}$  is the required total number of events.  $\Pr(event)$  in a survival study can be given by

$$\Pr(event) = \pi_A Pr_A(event) + \pi_B Pr_B(event), \quad (3.5)$$

where

$$\pi_A = \frac{R}{R + 1} \text{ and } \pi_B = \frac{1}{R + 1}, \quad (3.6)$$

are the proportions of patients who are randomized to experimental and control treatment group respectively and  $Pr_A(event)$  and  $Pr_B(event)$  are the probabilities of events in experimental and control arm respectively [120]. Freedman, 1982 [121]

provided an approximation of the probability of event for each treatment group assuming equal follow-up for all patients and thus simultaneous accrual for all patients whereas Schoenfeld, 1983 [122] provided a more exact approximation of the expected event rate as compared to Freedman's approximation. More precisely, according to Freedman's approximation,

$$Pr_i(event) \approx 1 - S_i(\tau) \quad (3.7)$$

and according to Schoenfeld's approximation,

$$Pr_i(event) \approx 1 - \{S_i(\tau) + 4S_i(T/2 + \tau) + S_i(T + \tau)\}/6, \quad (3.8)$$

where  $i$  denotes the corresponding treatment group (either experimental or control),  $\tau$  denotes the follow-up time and  $T$  the accrual period,  $T/2 + \tau$  denotes the median follow-up time and  $T + \tau$  denotes the total duration of the study. Another approximation of the probability of event could be

$$Pr_i(event) \approx 1 - S_i(T/2 + \tau) \quad (3.9)$$

considering that the survival probability can be approximated as the probability that a patient survives past the median follow-up time (i.e.  $T/2 + \tau$ ) [120].

The web-based interface of Zhao and Simon is composed of two options. If the first option is chosen, the treatment effects for assay-negative and assay-positive patients must be specified in order to evaluate the relative efficiency of enrichment and untargeted design, i.e. marker stratified design (see next section for further information) in which apart from the biomarker-positive patients, biomarker-negative patients are also included; if the second option is chosen, it is possible to account for error in the assaying of the study population, thus, both the treatment effects for target-negative and target-positive patients must be specified as well as the assay's sensitivity and specificity.

The sample size calculation using binary data is based on the formulae described by Simon and Maitournam [65, 111, 112] and again the two options offered

when assuming a time-to-event outcome are available, i.e. options both with and without accounting for error in biomarker status classification. When binary outcome is assumed and the allocation ratio is 1:1, the sample size of randomized patients required in each treatment arm (experimental and control) can be given as

$$N_{enrichment/arm} = 2\bar{p}_Q(1 - \bar{p}_Q) \left[ \frac{(z_{a/2} + z_\beta)}{(p_A^Q - p_B)} \right]^2, \quad (3.10)$$

where  $p_A^Q$  and  $p_B$  are the response probabilities in the experimental and control groups respectively,

$$\bar{p}_Q = \frac{p_A^Q + p_B}{2} \quad (3.11)$$

and  $z_{a/2}, z_\beta$  denote the upper  $a/2$ - and upper  $\beta$ -points respectively of a standard normal distribution where  $a$  and  $\beta$  denote the assumed type I error and type II error respectively. The response probability in the experimental group can be found by

$$p_A^Q = p_B + \delta_+, \quad (3.12)$$

where  $\delta_+$  denotes the improvement in response probability for biomarker-positive patients. Consequently, the total sample size of randomized patients will be

$$N_{enrichment} = 2N_{enrichment/arm} \quad (3.13)$$

For continuous response endpoints the aforementioned formula  $N_{enrichment/arm}$  changes to

$$N_{enrichment/arm} = \frac{2\sigma^2(z_{a/2} + z_\beta)^2}{(\mu_{A+} - \mu_{B+})^2}, \quad (3.14)$$

where  $\sigma^2$  denotes the anticipated common variance,  $\mu_{A+}$  and  $\mu_{B+}$  the mean responses for biomarker-positive patients in the experimental and control treatment arm respectively. These formulae are the standard formulae used for a traditional randomized trial.

In addition, if we want to account for error in the assaying of the study population, the number of patients to be randomized in each arm of the enrichment trial when using continuous response endpoints can be given by the following formula

$$N_{enrichment/arm} = 2\sigma^2(z_{\alpha/2} + z_{\beta})^2 \{\lambda_1[(1 - \omega)\zeta + \omega]\}^{-2} \quad (3.15)$$

where  $\omega$  measures the accuracy of the assay and corresponds to the PPV (positive predictive value of the assay, i.e. the proportion of patients who are assigned the biomarker-positive status according to the assay who are truly biomarker positive),  $\lambda_1$  is the treatment effect in the biomarker-positive patients and  $\zeta = \lambda_0/\lambda_1$  (where  $\lambda_0$  is the treatment effect in the biomarker-negative patients) [55].

Simon and Maitournam [65, 111, 112] considered that apart from the number of patients to be randomized, the number of patients needed to be screened should be also reported. Thus, they stated that the expected number of patients to be screened in the enrichment design is  $N_{enrichment}/k$  where  $k$  corresponds to the proportion of biomarker-positive patients. The online tool developed by Zhao and Simon provides both the number of patients to be screened and to be randomized.

**Statistical considerations:** Simon and Maitournam [65, 111, 112] undertook a simulation study, assuming a binary outcome, to compare power of the enrichment design with an untargeted design (i.e. marker stratified design, see next section for further information) in which all patients are randomized without measuring the biomarker. They concluded that the efficiency of the enrichment design relies both on the prevalence of the biomarker-positive patients and on the accuracy of the assay. In the situation where the assay cut-off point is not well established, there is a risk of severely compromising the power of the trial when using an enrichment design. However, assuming an accurate assay, if fewer than half of the entire study population are biomarker-positive and there is robust evidence that the experimental treatment does not benefit the biomarker-negative patients, the required number of randomized patients to allow sufficient power to detect a significant treatment effect

is much smaller in the enrichment design than in the untargeted trial design. On the other hand, in the latter situation a greater number of individuals would need to be screened when using the enrichment design, and accruing the required number of biomarker-positive patients could take a longer period of time. More precisely, Simon and Maitournam showed that an approximation of the ratio of the required number of patients to be randomized for the untargeted trial design as compared with the required number of patients randomized in the enrichment design when using binary outcome can be given by the following equation

$$\frac{N_{stratified}}{N_{enrichment}} \approx \frac{1}{\left[ k + (1 - k) \frac{\delta_-}{\delta_+} \right]^2} = \left[ \frac{\delta_+}{k\delta_+ + (1 - k)\delta_-} \right]^2, \quad (3.16)$$

where  $k$  denotes the proportion of biomarker-positive patients,  $\delta_-$  and  $\delta_+$  correspond to the treatment effectiveness (i.e. improvement in response probability) in biomarker-negative and biomarker-positive subgroups respectively. Consequently, in the situation where it is known that the novel treatment does not benefit the biomarker-negative patients at all, the ratio of the number of patients needed for randomization in the untargeted design relative to the number of patients required for the enrichment design is approximately

$$\frac{N_{stratified}}{N_{enrichment}} \approx \frac{1}{k^2}, \quad (3.17)$$

as  $\delta_- = 0$ . For example, if half of patients are biomarker-positive ( $k = 0.5$ ) then a quarter of those needed to be randomized to the untargeted design trial would need to be randomized to the enrichment design trial. In cases where the novel treatment is half as effective in biomarker-negative patients as in the biomarker-positive patients (i.e.  $\delta_-/\delta_+ = 1/2$ ), the aforementioned ratio changes to

$$\frac{N_{stratified}}{N_{enrichment}} \approx \frac{4}{(k+1)^2}. \quad (3.18)$$

### 3.2.3. Randomize-All Designs

---

Randomize-all designs (also named as all-comers/untargeted/unselected/non-targeted/simple randomization designs) allow the inclusion of the entire population as eligible for randomization. Consequently, the whole study population who meet the eligibility criteria, is randomly assigned to the different treatment groups (experimental and control treatment group) regardless of biomarker status. This design allows assessment of treatment benefit for the entire population irrespective of biomarker status whilst at the same time allowing for treatment benefit to be tested in the two biomarker-defined subgroups separately.

Generally, they are useful when we are uncertain about the benefit of the experimental treatment in the overall population versus the biomarker-defined subgroups, and the targeted treatment may benefit both biomarker-positive and biomarker-negative patients. Additionally, these designs are useful when the goal is to test the predictive ability of a biomarker, the assay reproducibility and accuracy is questionable, the turnaround time for biomarker assessment is long and the biomarker prevalence is high.

Randomize-all designs are composed of two main subtypes: the Marker-stratified designs and the Hybrid designs, which are discussed separately below.

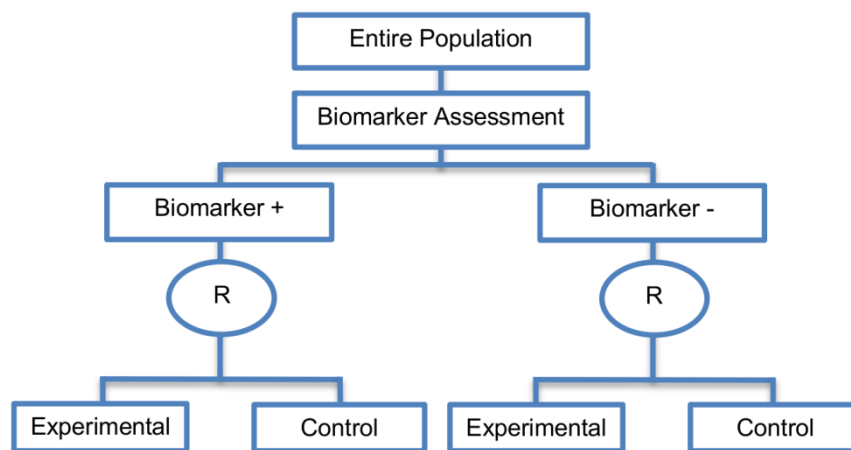
#### 3.2.3.1. Marker Stratified Designs

---

These designs (prospective validation Phase III trials) were identified in 45 papers (45%) of our review.

**Design:** An illustration of the design is shown in Figure 3.4. Individuals are stratified into biomarker-positive and biomarker-negative subgroups according to the results of the biomarker assessment and then they are randomized either to the experimental or to the control treatment group. The biomarker status in the Marker-

Stratified design acts as a stratification factor where stratification is used to ensure balance across treatment groups with regard to biomarkers. Only individuals with valid biomarker results enter the trial. Consequently, we have four treatment groups, i.e. biomarker-positive patients assigned to either the experimental treatment arm or the control treatment arm and biomarker-negative patients assigned to either the experimental treatment arm or the control treatment arm. Thus, we can assess the relationship between treatment effect and biomarker status.



**Figure 3.4.** Marker Stratified designs. “R” refers to randomization of patients.

**Utility:** When there is enough evidence that the experimental treatment is more effective in the positive biomarker-defined subgroup than in the negative biomarker-defined subgroup but there is no sufficient compelling data that the experimental treatment is of no benefit in biomarker-negative individuals, the marker stratified design can be used.

**Methodology:** Biomarker status is used to stratify the randomization, rather than to restrict eligibility. Marker-stratified designs can be conducted using two different testing plans; the so-called marker-by-treatment interaction with separate tests and marker-by-treatment interaction with interaction test. Both of these approaches involve conducting a single clinical trial, but treated as two independent ones for analysis purpose.

Marker-by-treatment interaction using separate test was referred to in 15 papers (15%) of our review [4, 11, 12, 15, 29, 42, 45, 53, 57, 60, 80, 82, 84, 87, 88] and is also referred to as 'separate randomization design' and 'separate by treatment interaction design'. This analysis plan is based on separate superiority tests in each biomarker-defined subgroup in order to detect the treatment efficacy in each subset. Two examples of actual trials which use this testing plan are the following: National Cancer Institute (NCI)-sponsored North Central Cancer Treatment Group Study N0975 [29] and the MARVEL trial [29].

The 'marker-by-treatment interaction design using separate tests' is a testing plan which determines whether the novel treatment is superior to the control treatment separately within each biomarker-defined subgroup. Consequently, the hypothesis to be tested, the calculation of the number of patients required for the trial, the estimation of the statistical power of the design and the randomization procedure of patients to different treatments are independent among the different subgroups [12]. The sample size of the trial should be calculated in such a way so as to yield adequate statistical power when testing whether the experimental treatment is superior to the control treatment separately in the two biomarker-defined subgroups. Hence, this approach is not widely used due to the required large sample size as essentially two separate trials are being conducted, and in addition interaction tests increase sample size requirements even further. Another limitation of this approach is that when multiple biomarker-defined subsets and treatments are to be investigated, it is difficult to implement in practice.

The 'marker-by-treatment interaction using interaction test' uses a test for interaction between the biomarker status and treatment assignment and was identified in 12 papers (12%) of our review [4, 12, 15, 42, 53, 57, 60, 82, 84, 87, 88, 94]. A marker stratified design which uses this testing plan is also referred to in the literature as an 'interaction design' or 'genomic signature stratified design'. First, a formal statistical test for interaction between biomarker status and treatment assignment is undertaken. If this interaction is not significant, then the study is continued by testing the different treatments overall at a two-sided significance level



of 0.05, otherwise, the treatments are compared within each biomarker-defined subpopulation at a two-sided 0.05 significance level (i.e. the same as in the marker-by-treatment interaction design using separate tests). The sample size for this second testing plan is calculated with reference to the treatment effect in the entire study population. Therefore, it might not provide sufficient power for detecting the treatment effect in each biomarker defined-subset individually. More precisely, if the sample size is calculated for the overall analysis and the proportion of the biomarker-defined subpopulation which responds to the novel treatment is very small, the statistical power for the subgroup analysis may be inadequate. In addition, when several biomarker-defined subpopulations and treatments are to be investigated, this strategy is not easy to be implemented.

For the case of binary outcomes, Eng, 2014 [92] provided the formula for the required sample size to power the biomarker-positive and biomarker-negative patients separately. It is assumed that  $Y$  is a binary variable which corresponds to a patient's response to their randomly tailored treatment and  $P(Y|Trt = i, M = j) = r_{ij}$  where  $i$  corresponds to either the experimental or control treatment and  $j$  corresponds to either the biomarker-positive patients or the biomarker-negative patients. Hence,

$$r_{ij} = \beta_0 + \beta_A I(Trt = A) + \beta_+ I(M = M^+) + \beta_l I(Trt = A, M = M^+), \quad (3.19)$$

where  $\beta_0$  denotes a baseline effect,  $\beta_A$  denotes the added effect of the experimental treatment,  $\beta_+$  denotes the biomarker-positive effect and  $\beta_l$  denotes the nonadditive effect. Consequently, the proposed formula for the required sample size can be given by

$$N_{stratified} = 2(z_a + z_{1-\beta})^2 \left\{ \frac{r_{A+}(1 - r_{A+}) + r_{B+}(1 - r_{B+})}{(\beta_A + \beta_l)^2} + \frac{r_{A-}(1 - r_{A-}) + r_{B-}(1 - r_{B-})}{(\beta_A)^2} \right\}, \quad (3.20)$$

where  $a$  corresponds to the target level,  $1 - \beta$  corresponds to the power. Also,  $r_{A+}, r_{B+}$  are the assumed response rates of biomarker-positive patients receiving the

experimental and the control treatment respectively. Additionally,  $r_{A-}, r_{B-}$  are the assumed response rates of biomarker-negative patients receiving the experimental and the control treatment respectively.

Mandrekar and Sargent, 2009 [31] provide a formula to calculate the required number of events when the trial has a survival outcome with 1:1 randomization to treatment arms, i.e.

$$D_{stratified} = \frac{4(z_{a/2} + z_{\beta})^2}{\left[\log\left(\frac{m_{A+}}{m_{B+}}\right)\right]^2} + \frac{4(z_{a/2} + z_{\beta})^2}{\left[\log\left(\frac{m_{A-}}{m_{B-}}\right)\right]^2}, \quad (3.21)$$

where  $m_{A+}, m_{A-}, m_{B+}, m_{B-}$ , indicate the median overall survival for biomarker-positive and biomarker-negative patients receiving control and experimental treatment, respectively and

$$\theta_1 = \frac{m_{A+}}{m_{B+}} = HR_{biom^+}, \quad (3.22)$$

$$\theta_2 = \frac{m_{A-}}{m_{B-}} = HR_{biom^-}, \quad (3.23)$$

correspond to the hazard ratios of biomarker-positive and biomarker-negative subgroups and  $z_{a/2}, z_{\beta}$  denote the upper  $a/2$ - and upper  $\beta$ -points respectively of a standard normal distribution where  $a$  and  $\beta$  denote the assumed type I error and type II error respectively. More precisely, the total number of events is the sum of the required number of events for the biomarker-negative and biomarker-positive subpopulation. Freidlin et al., 2010 [61] stated that the required number of events in order to compare the experimental to the control treatment among the biomarker-positive patients for detecting a given effect size in this biomarker-positive subpopulation is identical to the number of events needed by an enrichment design (i.e.  $D_{enrichment}$ ).

Another potential formula for the required total number of events when 1:1 randomization to treatment arms is assumed is given by

$$D_{stratified} = \frac{4(z_{\alpha/2} + z_{\beta})^2}{[k\log(\theta_1) + (1 - k)\log(\theta_2)]^2}. \quad (3.24)$$

Although the formula proposed by Mandrekar and Sargent, 2009 [31] achieves a specific power  $(1 - \beta)$  for each biomarker-defined subgroup separately, the aforementioned formula proposed in the book of Harrington, 2012 [114] aims to reach a power  $(1 - \beta)$  for the overall population to test the null hypothesis of no treatment effect in the entire population. According to Harrington, 2012 the required total number of patients to be entered to a stratified trial can be given by

$$N_{stratified} = \frac{4(z_{\alpha/2} + z_{\beta})^2}{\left\{ \frac{[kPr_{(+)}(event)\log(\theta_1) + (1 - k)Pr_{(-)}(event)\log(\theta_2)]}{\sqrt{kPr_{(+)}(event) + (1 - k)Pr_{(-)}(event)}} \right\}^2}, \quad (3.25)$$

where  $Pr_{(+)}(event)$ ,  $Pr_{(-)}(event)$  are the probabilities of an event in biomarker-positive subset and biomarker-negative subset respectively. If we divide the required total number of events for the enrichment design by the aforementioned formula for the required total number of events for the stratified design, we can get the following approximation of the ratio

$$\begin{aligned} \frac{D_{stratified}}{D_{enrichment}} &= \frac{[\log(\theta_1)]^2}{[k\log(\theta_1) + (1 - k)\log(\theta_2)]^2} \\ &= \frac{1}{\left[ k + (1 - k) \frac{\log(\theta_2)}{\log(\theta_1)} \right]^2}. \end{aligned} \quad (3.26)$$

Further, Zhao and Simon [19, 28, 53, 57, 60] have developed an online tool for the calculation of sample size for biomarker stratified randomized designs with binary or time-to-event endpoints which is available online at the following web site <http://brb.nci.nih.gov/brb/samplesize/sdpap.html> [115]. More precisely, the sample size for both binary and time-to-event endpoints can be performed with three different analysis plans; A, B and C. Before choosing one of these analysis plans in the web site, for binary endpoints we need to specify the probability of treatment response in the control arm as well as the proportion of biomarker-positive patients.

For survival endpoints, the hazard ratio of biomarker-positive patients versus the biomarker-negative control patients which corresponds to the hazard ratio of prognostic effect as well as the proportion of biomarker-positive patients must be specified.

Analysis plan A is performed when there is confidence that an overall treatment effect exists. It determines the sample size on the basis of first of all comparing the experimental treatment to the control treatment in the entire randomized population at a reduced two-sided significance level  $\alpha < 0.05$ . If the overall test is not significant, then the experimental treatment is compared to the control treatment in the biomarker-positive patients using the significance level  $0.05 - \alpha$ . Analysis Plan A is similar to the 'Biomarker-positive and overall strategies design' with fall-back analysis described later in this chapter; the difference lies in this in terms of the significance levels they have used. In order for the sample size to be estimated, the anticipated overall effect estimate, reduced two-sided significance level and power for the overall test need to be specified.

Analysis plan B is performed when there is confidence that there is a treatment effect in the biomarker-positive subpopulation. It determines the sample size on the basis of first of all comparing the experimental treatment to the control treatment in the biomarker-positive subgroup at a two-sided significance level of  $\alpha = 0.05$  level. If the treatment effect is found to be significant at this 0.05 level, then treatment effect is evaluated in the biomarker-negative subgroup again at a two-sided significance level of 0.05 level. This analysis plan is identical to the 'Sequential subgroup-specific design' described later in this chapter. In order for the sample size to be estimated, apart from the fixed significance level set to 0.05, the anticipated effect estimate in the biomarker-positive subpopulation and power need to be specified.

Analysis plan C first tests whether there is a statistically significant interaction between treatment and biomarker [60]. If the interaction is not significant, then the treatments are compared in the overall study population at a two-sided significance level 0.05. Otherwise, the treatments are compared within the two biomarker

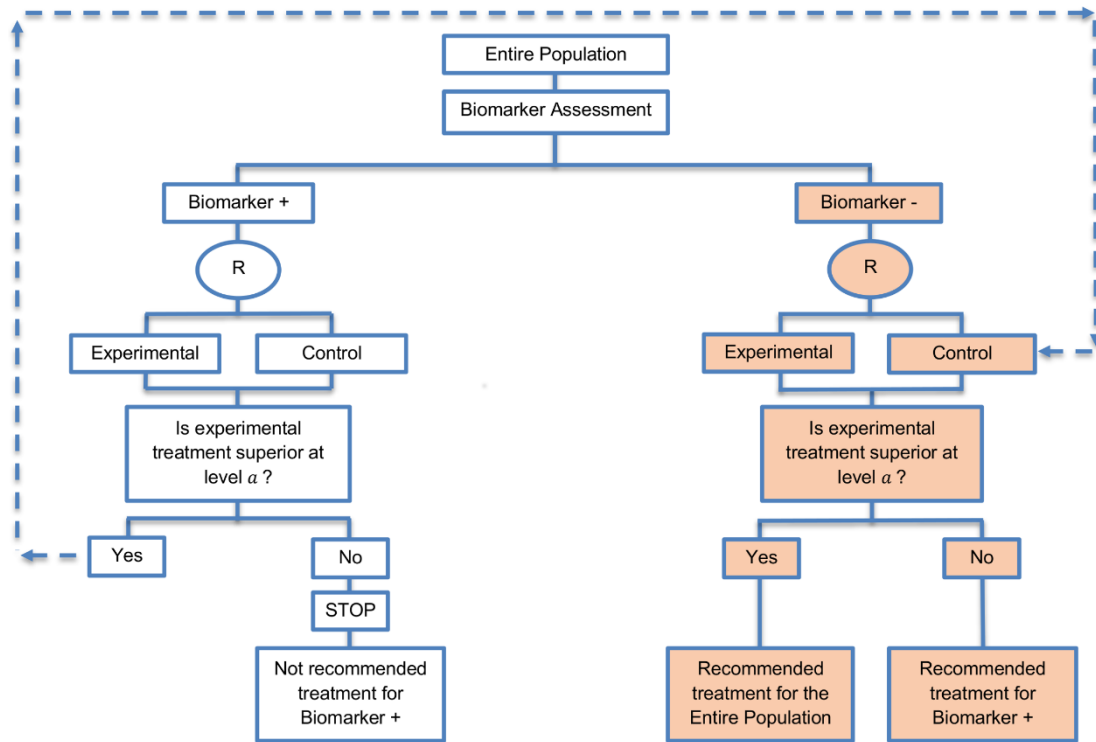
subgroups separately at a two-sided 0.05 significance level for each subgroup. Analysis Plan C follows either the ‘marker-by-treatment interaction process with interaction or the separate test process’ described above. In order for the sample size to be estimated, the anticipated treatment effect in the overall study population, the one-sided significance level for interaction test and the power for testing the treatment effect in the overall population need to be specified.

In marker stratified designs, three designs can be included which differ in terms of their statistical testing strategies, i.e. (i) Subgroup-specific designs (i.e. sequential subgroup-specific design, parallel subgroup-specific design); (ii) Biomarker-positive and overall strategies (i.e. biomarker-positive and overall strategies with parallel assessment, biomarker-positive and overall strategies with sequential assessment, biomarker-positive and overall strategies with fall-back analysis); (iii) Marker sequential test design (MaST) and they are discussed in the following sections.

**Statistical considerations:** Despite the fact that the marker stratified designs allow testing the treatment effect not only in the entire population but also in each biomarker-defined subpopulation, they might not be feasible when the prevalence of biomarker is low. Another limitation of such designs is that they might require a large sample size where several treatments and biomarkers are investigated in the study.

**Subgroup-Specific designs:** This strategy is an approach to analyze a biomarker-stratified trial. It is composed of two types; ‘Sequential Subgroup-Specific design’ and ‘Parallel Subgroup Specific design’. Both biomarker-positive and biomarker-negative subgroups can be tested in a sequential or in a parallel way. With the parallel way, we can assess simultaneously both biomarker-positive and biomarker-negative patients, whereas, with the sequential way we perform first the assessment of biomarker-positive patients and if the result is positive then we continue with the biomarker-negative patients.

**Sequential Subgroup-Specific design:** This approach was referred to in 11 papers (11%) of our review. Figure 3.5 graphically represents this approach.



**Figure 3.5.** Sequential Subgroup-Specific design. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. Different stages refer to the analysis and not to the trial design.

**Design:** The sequential testing procedure uses the assumption that it is unlikely that the new treatment will be effective in the biomarker-negative patients unless it is effective in the biomarker-positive patients. First, the treatment effect is tested in the biomarker-positive subpopulation using the overall two-sided significance level  $\alpha = 0.05$  (Type I error); if this test is significant then the treatment effect is tested in the biomarker-negative subgroup using the same level of significance  $\alpha$ .

**Utility:** Its use is recommended when there is compelling evidence that biomarker-positive individuals benefit more from the experimental treatment than the biomarker-negative patients. More precisely, it is appropriate when it is not expected for the novel treatment to be effective in biomarker-negative patients unless it is beneficial for the biomarker-positive patients.

**Methodology:** As this subgroup-specific design follows a sequential assessment and thus the design is composed of two stages, the sample size calculation

is also staged. For binary outcome the required number of biomarker-positive patients is the same as for the enrichment design, i.e.

$$N_{\text{Sequential subgroup-specific}}^+ = N_{\text{enrichment}} \quad (3.27)$$

As Simon, 2008 [60] stated, the total number of patients will be approximately

$$N_{\text{Sequential subgroup-specific}} = \frac{N_{\text{enrichment}}}{k} \quad (3.28)$$

where  $k$  is the proportion of biomarker-positive patients and the number of biomarker-negative patients will be approximately

$$N_{\text{Sequential subgroup-specific}}^- = \frac{(1 - k)N_{\text{enrichment}}}{k}. \quad (3.29)$$

For the conduct of this design, it is important to ensure that there is also an adequate number of biomarker-negative patients for analysis purposes. For time-to-event outcomes, the required number of events for biomarker-positive patients is the same with the required number of events in the enrichment design, i.e.

$$D_{\text{Sequential subgroup-specific}}^+ = D_{\text{enrichment}}. \quad (3.30)$$

At the time that there are  $D_{\text{enrichment}}$  patients, the required number of events among biomarker-negative patients in terms of that among biomarker-positive patients ( $D_{\text{enrichment}}$ ) is given by

$$D_{\text{Sequential subgroup-specific}}^- = D_{\text{enrichment}} \left( \frac{\lambda_-}{\lambda_+} \right) \left( \frac{1 - k}{k} \right), \quad (3.31)$$

where  $\lambda_-$ ,  $\lambda_+$  are the event rates in biomarker-negative and biomarker-positive control subsets at the time when there are  $D_{\text{enrichment}}$  events in the biomarker-positive subgroup [60].

The significance levels  $\alpha$  can also be considered as one-sided significance levels in situations where our alternative hypothesis is not that there is just a treatment

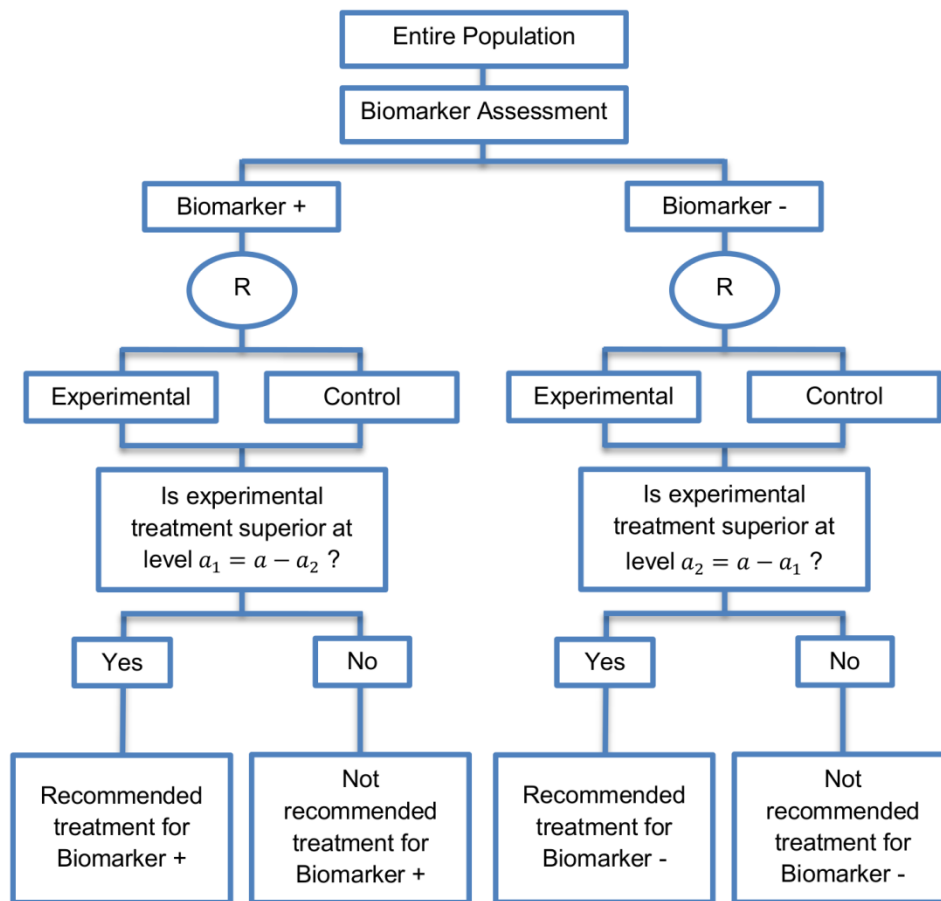
effect but that the treatment benefit in the experimental group is greater than that of the control group.

**Statistical considerations:** This strategy preserves the overall type I error rate  $\alpha$  but requires a smaller number of positive patients as compared to the second type of subgroup-specific design, the so-called parallel subgroup-specific design (see below). Furthermore, it enables the identification of treatment efficacy in the biomarker-positive and biomarker-negative subpopulations separately. However, it yields low power when there is homogeneity of treatment effect across the different biomarker-defined subpopulations. Furthermore, in case that test for treatment effect among biomarker-negative patients is not statistically significant, an “exploratory” analysis on the biomarker-negative subgroup might be considered.

**Parallel Subgroup-Specific design:** This design was identified in three papers (3%) of our review.

**Design:** Parallel subgroup-specific design (Phase III), also referred to as a Phase III Biomarker-Stratified design evaluates treatment effects separately in the positive biomarker-defined subgroup and in the negative biomarker-defined subgroup simultaneously. A graphical illustration of this strategy is given in Figure 3.6.





**Figure 3.6.** Parallel Subgroup-Specific design. “R” refers to randomization of patients.

**Utility:** It is appropriate when the aim of the study is to give treatment recommendations for each biomarker-defined subgroup separately at the same time.

**Methodology:** In order to control the overall type I error rate of the design at the overall level of significance  $\alpha$  (Type I error) it is required to allocate this overall  $\alpha$  between the test for the biomarker-positive subgroup and the test for the biomarker-negative subgroup using the Bonferroni correction method [123] for multiple testing; e.g., if we choose the value of 0.025 for the global significance level  $\alpha$ , then we could choose the values of  $\alpha_1 = 0.010$  and  $\alpha_2 = 0.015$  for testing the biomarker-negative and biomarker-positive subgroups respectively. This trial design is powered in such a way so as to detect the treatment effect in each biomarker-defined subgroup separately. A higher portion of the type I error rate can be given for the test within the biomarker-positive subgroup in order to maximize the power of the trial to identify the treatment effect in this subpopulation.

As in the sequential subgroup-specific design, the probability of rejecting either the null hypothesis of no treatment effect in the biomarker-positive subset or in the biomarker-negative effect under the global null hypothesis is less than or equal to the overall type I error rate  $\alpha$ . Additionally, the probability of rejecting the null hypothesis of no treatment effect in the biomarker-negative subpopulation when the treatment benefit is only restricted to biomarker-positive patients is less than or equal to  $\alpha$ . The significance levels  $\alpha$  can be considered as one-sided or two-sided significance levels.

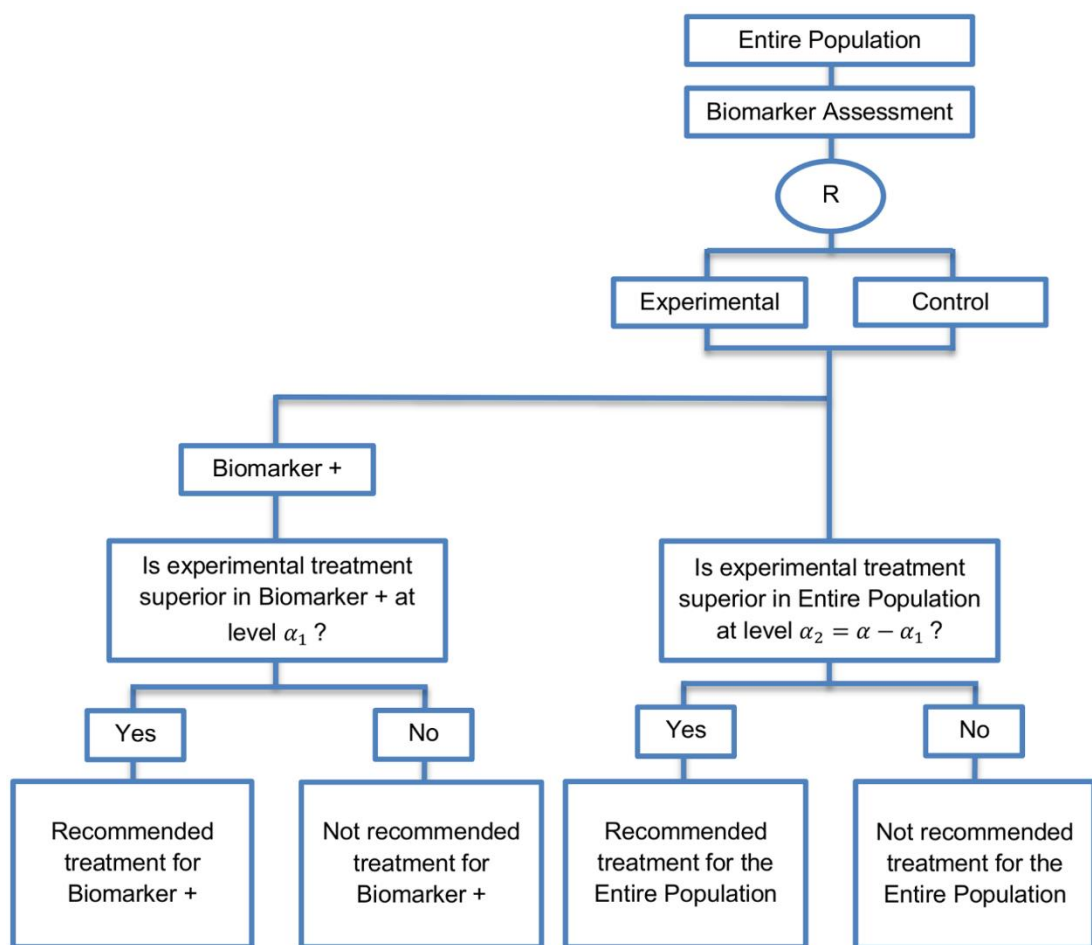
**Statistical considerations:** With this approach, in case that the overall level of significance  $\alpha$  is equal in both subgroup-specific designs, it is more difficult to achieve statistical significance in the biomarker-positive subgroup as compared to the sequential subgroup-specific design due to the allocation of the overall significance level between the two biomarker-defined subgroup tests.

**Biomarker-positive and overall strategies:** This design provides an alternative strategy to analyzing a biomarker-stratified design. It is an indirect way of evaluating both biomarker and treatment by testing the treatment effect in the entire study population and in the biomarker-positive subgroup separately. Three approaches are included in the biomarker-positive and overall strategies; the parallel assessment, the sequential assessment and the fall-back design (see below).

Despite the fact that the biomarker-positive subgroup and overall strategy design allows the treatment effect to be tested in the biomarker-positive subpopulation and provides good statistical power when the treatment effect is homogeneous across subgroups, this design is usually considered problematic and its use is not often recommended. More precisely, a major concern is that when the benefit of the novel treatment is limited to the biomarker-positive patients, it is possible that the design might lead to a wrong recommendation of treatment for the biomarker-negative patients. This might happen because when there is no treatment effect in the biomarker-negative subgroup, there might be an observed effect in the entire population due to the potentially large effect in the biomarker-positive

patients. This concern is particularly pronounced in the sequential version of the design, which first tests the biomarker-positive subgroup and then, if it is positive, it tests the overall population.

**Biomarker-positive and overall strategies with parallel assessment:** This approach was identified in eight papers (8%) of our review. Figure 3.7 graphically represents this strategy. In the parallel version, we test both the overall population and biomarker-positive subgroup simultaneously.



**Figure 3.7.** Biomarker-positive and overall strategies with parallel assessment. “R” refers to randomization of patients.

**Design:** In this approach the treatment effect is tested in both the entire study population and in the biomarker-positive patients while controlling the type I error by allocating the overall significance level  $\alpha$  between the two tests. The significance level  $\alpha$  can be considered as one-sided or two-sided.

**Utility:** The parallel version is recommended when the aim of the study is to assess the treatment effect in both the overall study population and in the biomarker-positive subgroup but not in the biomarker-negative subgroup.

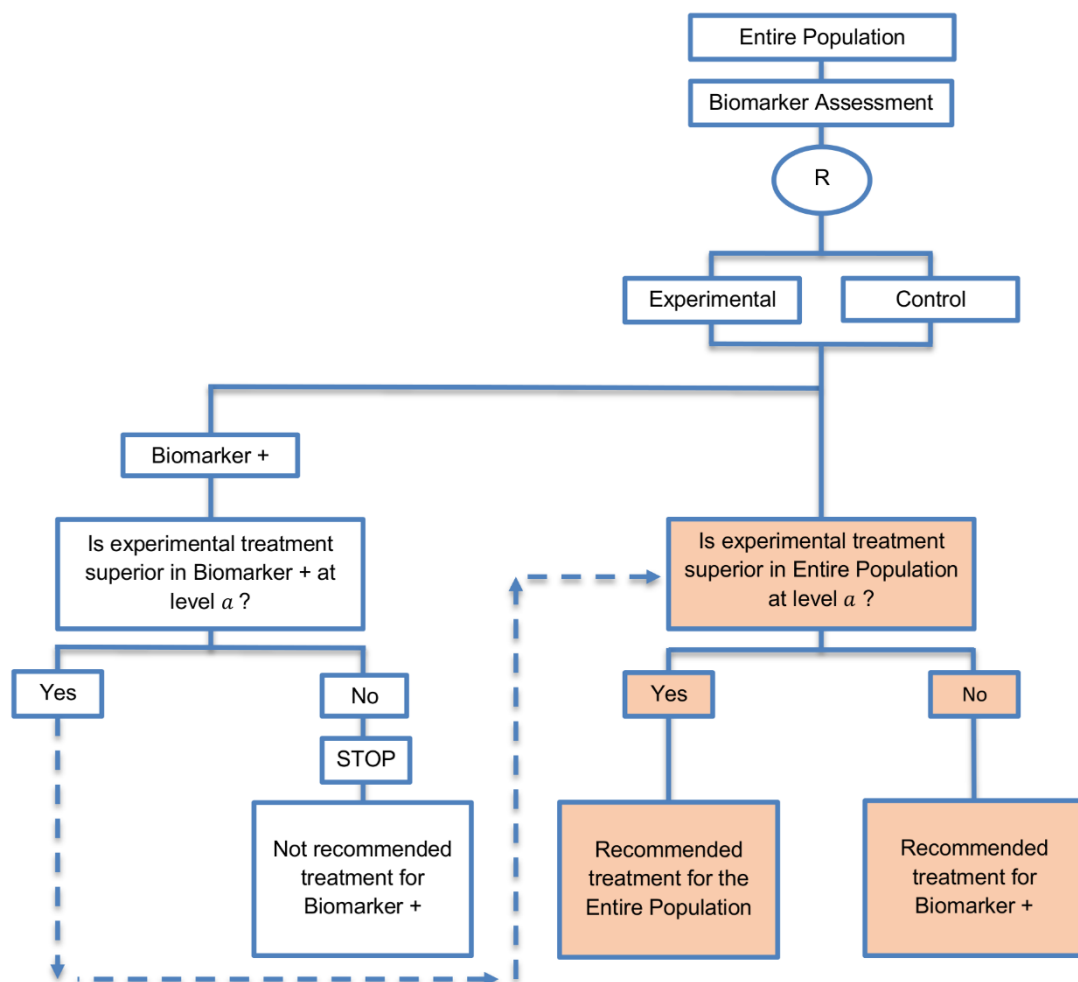
**Methodology:** If there is significant confidence that the biomarker is predictive, the sample size estimation is aimed at having a sufficient number of biomarker-positive individuals to enable the treatment effect in the biomarker positive subgroup to be detected. On the other hand, if there is no confidence in the predictive value of the biomarker, the sample size estimation is aimed at having a sufficient number of patients to detect a treatment effect in the overall study population [14].

**Statistical considerations:** This design has the ability to control the probability of rejecting the null hypothesis of no treatment effect either in the biomarker-positive population or in the biomarker-negative population under the global null hypothesis of no treatment effect in the entire population at the overall significance level  $\alpha$ . However, it cannot control the probability of rejecting the null hypothesis of no treatment effect in the biomarker-negative subset when the treatment benefit is restricted to biomarker-positive patients. Consequently, there is high risk of inappropriately recommending the experimental treatment for biomarker-negative patients.

When the experimental treatment is compared to the control treatment within the overall population and the overall treatment effect is significant, then the test has high statistical power. If we are testing only the biomarker-positive subgroup and the treatment effect in this subgroup is significant, the statistical power is again high. This prospective subset analysis plan is based on testing both the overall study population and the biomarker-positive subgroup using significance levels, which are chosen in such a way that the overall significance level is equal or less than  $\alpha$  (type I error). An easy way is to split  $\alpha$  in such a way that the significance level for the entire population and the significance level for the biomarker-positive subset equals to overall significance level  $\alpha$  (typically  $\alpha = 0.05$ ). For example, the SATURN trial (NCT00556712) [96] which employs a prospective subset strategy used the value of

0.03 as level of significance to test the treatment effect in the entire population and the value of 0.02 to test the treatment effect in the biomarker-positive subset; therefore, the overall level of significance was preserved at 0.05. The approach can be overly conservative as in the SATURN trial because of the correlation between the global and subgroup test. Other approaches [98, 124-127] have been proposed for adjusting the level of significance of both tests in a more accurate and less conservative way.

**Biomarker-positive and overall strategies with sequential assessment:** This approach was referred to in 11 papers (11%) of our review. A graphical illustration of this approach is shown in Figure 3.8.



**Figure 3.8.** Biomarker-positive and overall strategies with sequential assessment. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. Different stages refer to the analysis and not to the trial design.

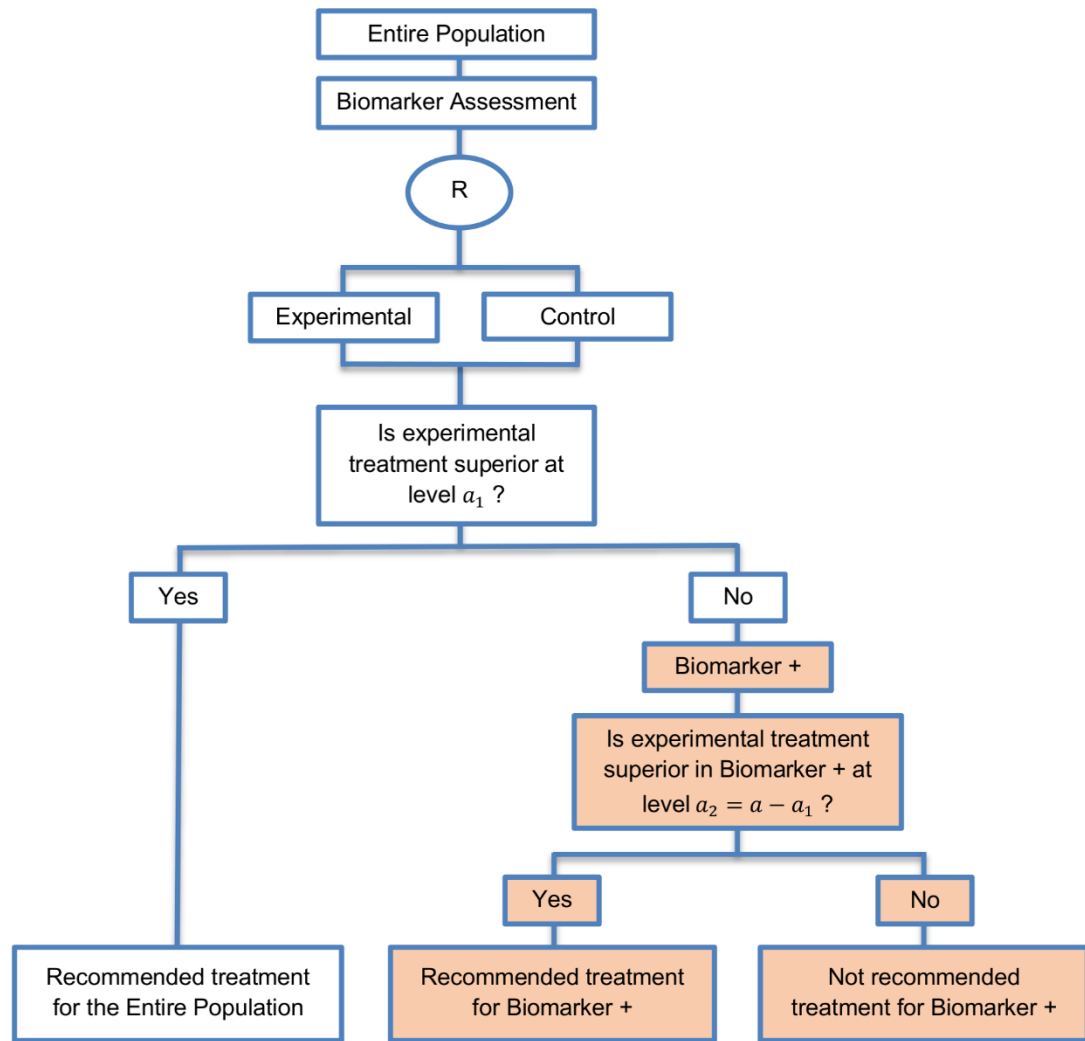
**Design:** In this sequential version of the biomarker-positive and overall strategies, we first test the biomarker-positive subgroup using the significance level  $\alpha$ ; if the test is significant, then we test the treatment effect in the overall population using the same  $\alpha$  level. The significance levels  $\alpha$  can be considered as one-sided or two-sided significance levels.

**Utility:** The sequential version might be useful in cases where the experimental treatment is expected to be effective in the overall study population.

**Methodology:** As this design comprises two sequential stages, it follows that the sample size calculation should also be staged. At the first stage, the standard formula for a traditional randomized trial can be used for the biomarker-positive subgroup using the significance level  $\alpha$  to estimate the treatment effect in that subset. More precisely, the formula used in the enrichment design for the required total number of events or the required number of patients can be used at the first stage of this design. At the second stage, the sample size must be adjusted in order to yield appropriate power for the entire population.

**Statistical considerations:** As in the parallel version of this designs, this strategy does not allow for identification of treatment efficacy in the biomarker-negative subgroup and despite the fact that it can control the overall type I error  $\alpha$  it cannot control the probability of rejecting the null hypothesis of no treatment effect in the biomarker-negative subset when the treatment benefit is restricted to biomarker-positive patients. Consequently, for this design also there is high risk of inappropriately recommending the novel treatment for biomarker-negative patients.

**Biomarker-positive and overall strategies with fall-back analysis:** This strategy was identified in 15 papers (15%) of our review. It evaluates both the treatment effect in the overall study population and in the biomarker-positive subgroup sequentially. Figure 3.9 graphically represents this strategy.



**Figure 3.9.** Biomarker-positive and overall strategies with fall-back analysis. “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. Different stages refer to the analysis and not to the trial design.

**Design:** In the fall-back design, we first test the overall population using the reduced significance level  $a_1$  and if the test is significant, we consider that the novel treatment is effective in the overall population; however, if the result is not significant then we test the treatment effect in the biomarker-positive subgroup using the level of significance  $a_2 = a - a_1$ , where  $a$  is the overall significance level (Type I error rate). The significance levels  $a$  can be considered as one-sided or two-sided significance levels. **Utility:** This approach is recommended when there is insufficient confidence in the predictive value of the biomarker and that the novel treatment is believed to be effective in all individuals (i.e. the rationale for the biomarker is weak). This design can be used in order to avoid the possibility of missing an important treatment effect

in the biomarker-positive patients (with insufficient benefit in the biomarker-negative subgroup).

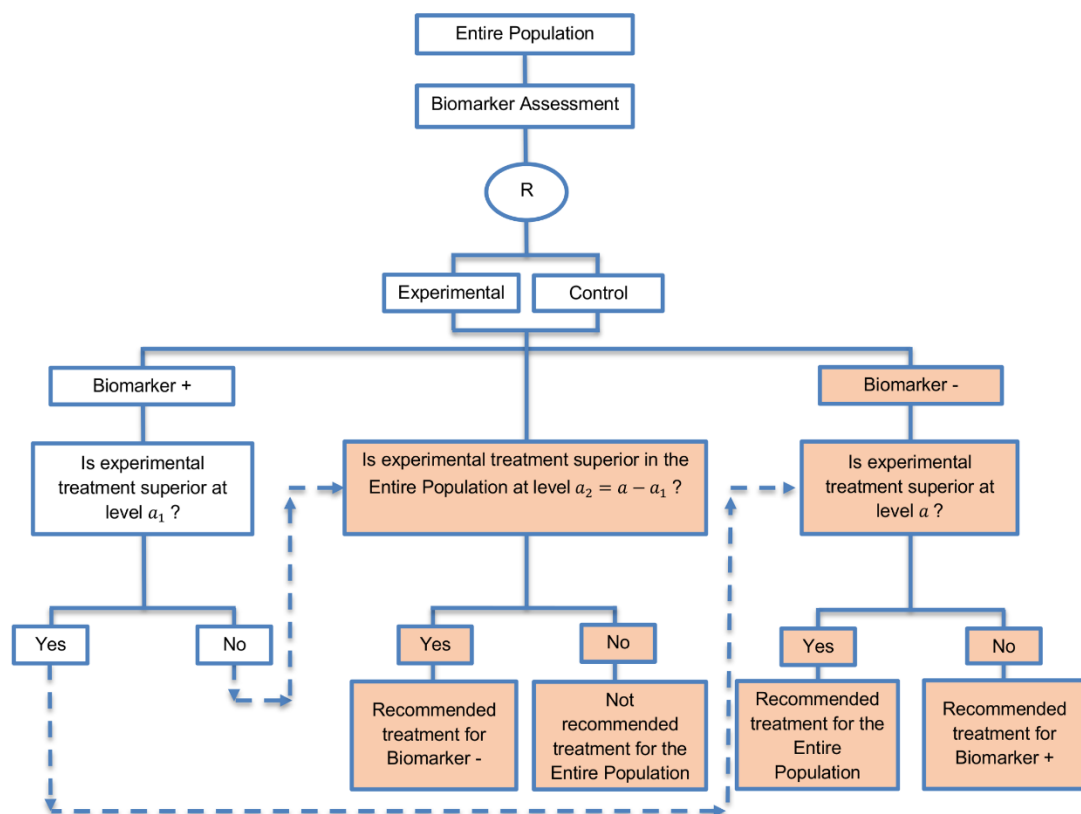
**Methodology:** The sample size should be set in such a way so as to yield adequate power for the overall test at the reduced significance level  $\alpha_1$  and for the potential biomarker positive subgroup analysis at significance level  $\alpha - \alpha_1$  [60]. The fall-back version is identical to the parallel version of biomarker-positive and overall strategies in terms of sample sizes and study outcomes, however the difference between these approaches is that the fall-back strategy is useful in settings where a biomarker will be assessed only if the overall population benefit is not promising [14]. This strategy can test the treatment effectiveness in biomarker-positive patients even if there is no detected benefit of the novel treatment in the overall population. However, it does not evaluate clearly the treatment benefit in the biomarker-negative subpopulation.

**Statistical considerations:** As the two aforementioned biomarker-positive and overall designs, this strategy can again control the overall type I error  $\alpha$  but it cannot control the probability of rejecting the null hypothesis of no treatment effect in the biomarker-negative subgroup when the treatment benefit is restricted to biomarker-positive patients. Consequently, there is high risk of inappropriately recommending the novel treatment for biomarker-negative patients. Song et al., 2007 [128] and George, 2008 [1] have discussed refinement of the significance levels associated with this design, which takes into account the correlation between the test for overall treatment effect and the test for the biomarker-positive treatment effect [60]. Additionally, a recent paper by Choai et al., 2015 [97] proposes a bias-corrected estimation method for treatment effects for the all-comers randomized clinical trials with a predictive biomarker which incorporate the fall-back analysis. For Choai et al., 2015 [97] the terminology “all-comers randomized clinical trials” is referred to the “Biomarker-positive and overall strategies with fall-back analysis”. More precisely, as this study design has an adaptive nature and is composed of two stages, a bias is possible to arise in the treatment effect estimation in the biomarker-positive subset when the first stage of the trial yields an overall result which is not significant and



thus fails to demonstrate a treatment efficacy in the entire population. For this reason, Choai et al. ,2015 [97], formulate a bias function using polynomials in order to take into account the possibility of failing to demonstrate overall treatment efficacy during the first stage of the trial.

**Marker Sequential test design (MaST):** This design was identified in four papers (4%) of our review and while controlling the appropriate type I error rates, it evaluates not only the biomarker-positive and biomarker-negative subgroups but also the entire population sequentially to limit the assessment of treatment effect in the overall population when it seems that the biomarker-positive subgroup does not benefit from the novel treatment. A graphical illustration of this approach is given in Figure 3.10.



**Figure 3.10.** Marker Sequential test design (MaST). “R” refers to randomization of patients. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial.

**Design:** In this design which owns an adaptive nature, first, the biomarker-positive subgroup is tested at a reduced level  $a$  in  $[0, a]$  and if the result is significant,

then the biomarker-negative subgroup is tested at the global significance level  $\alpha$ . Otherwise, if the result is not significant, then the overall population is tested at level  $\alpha_2 = \alpha - \alpha_1$  in order to make a treatment recommendation for the biomarker-negative patients.

**Utility:** It is generally recommended when robust evidence is available regarding a biomarker and there is prior evidence showing that the novel treatment is more beneficial for the biomarker-positive patients as compared to the biomarker-negative patients. Additionally, it is appropriate when we can assume that the treatment will not be beneficial for the biomarker-negative subgroup unless it is effective for the biomarker-positive subgroup. Additionally, the marker sequential test design is considered as an alternative to the sequential subgroup-specific design when the aim is to consider the treatment effect not only in biomarker-positive but also in the biomarker-negative patients.

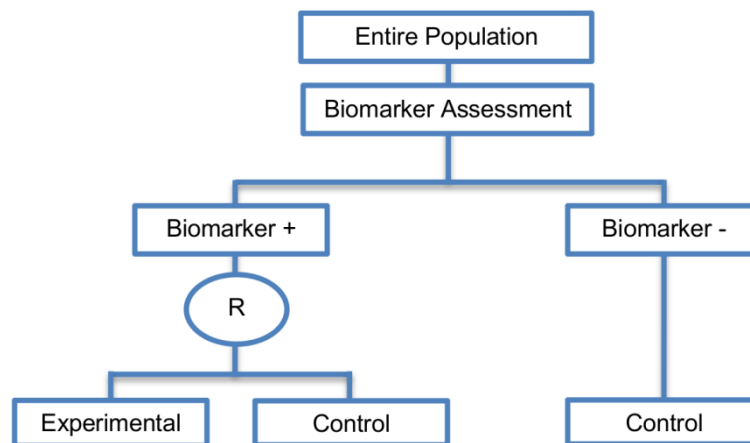
**Methodology:** Freidlin et al., 2014 [69] recommended using the value of 0.022 for the reduced significance level  $\alpha_1$  in order to control the type I error rate for biomarker-negative patients at the global significance level  $\alpha = 0.025$  and the value of 0.04 for the reduced significance level  $\alpha_1$  in order to control the type I error rate for biomarker-negative patients at the global significance level  $\alpha = 0.05$ .

Regarding the sample size for such a design where there is prior evidence indicating strong predictive ability of the biomarker, a standard sample size calculation (i.e. the same sample size calculation as for the enrichment designs) can be used for biomarker-positive subpopulation or alternatively, researchers can use the sample size calculation used for the sequential subgroup-specific design. However, in order to have sufficient number of biomarker-positive patients to detect treatment effectiveness in that particular biomarker-defined subset and consequently to reach the desired power, the sample size should be calculated using the reduced level  $\alpha_1$   $[0, \alpha]$  instead of the global significance level  $\alpha$  which is used in the sample size formulae of the enrichment and sequential subgroup-specific designs. This will result in a small increase in the number of patients as compared to the enrichment

and sequential subgroup-specific designs. Otherwise, if the reduced significance level  $\alpha_1$  is not used, this would yield minor loss of power.

**Statistical consideration:** Freidlin et al., 2014 [69] performed a comparison between the MaST and the sequential subgroup-specific design through a simulation study and concluded that the marker sequential design yields higher power in cases where the treatment effect is homogeneous across biomarker-defined subgroups. Additionally, with this approach, the power is preserved in situations where the experimental treatment is effective only for the biomarker-positive patients. Furthermore, in situations where biomarker status is not available for a portion of patients included in the trial, the marker sequential test design can either exclude these patients or include them in the global test, whereas, the proposed subgroup-specific designs do not consider inclusion of these patients in the analyses. If researchers decide to exclude patients with unavailable biomarker status from the study when using a MaST design, no statistical adjustment is required. On the other hand, if the inclusion of this study population is chosen, then this can result in inflation of the type I error rate for the biomarker-negative subpopulation above the global significance level  $\alpha$  due to the modification of correlation structure between the biomarker-defined subgroup tests and global test. In addition, while both MaST and subgroup-specific designs have the ability to control the probability of incorrectly rejecting the null hypothesis of no treatment effect in the biomarker-negative patients at the significance level  $\alpha$  when the experimental treatment does not work in either biomarker-defined subgroup, the sequential subgroup-specific approach typically has a smaller probability of incorrectly rejecting the null hypothesis of no treatment effect in the biomarker-negative subset (when the null hypothesis is true) as compared to the MaST design, especially under the global null hypothesis of no treatment effect in the entire population; the probability of incorrectly rejecting the null hypothesis of no treatment effect in the biomarker-negative patients depends on the choice of  $\alpha_1$ . This conservativeness of sequential subgroup-specific design, which is due to its sequential nature, makes the MaST design advantageous [69].

Hybrid designs (Phase III) were identified in 14 papers (14%) of our review and they can be included in the all-comers designs, where the entire population is firstly screened for biomarker status and all individuals enter the trial. A graphical illustration of this design is given in Figure 3.11.



**Figure 3.11.** Hybrid design. “R” refers to randomization of patients.

**Design:** In this approach, only the biomarker-positive patients are randomly assigned to either the experimental treatment group or to the control treatment group whereas the biomarker-negative patients receive the control treatment. These designs were first defined by Mandrekar and Sargent [30, 31]. The difference compared with the enrichment designs is that the biomarker-negative patients are not excluded from the study.

**Utility:** Hybrid designs can be used when there is compelling prior evidence which shows detrimental effect of the experimental treatment for a specific biomarker-defined subgroup (i.e. biomarker-negative subgroup) or some indication of its possible excessive toxicity in that subgroup, thus making it unethical to randomize the patients within this population to the experimental treatment.

**Methodology:** Similar to the enrichment design, hybrid designs are powered to identify the treatment effect only in the biomarker-defined subgroup which is randomly assigned to the experimental or control treatment groups. Consequently,

the same formula used for the required number of patients or events for the enrichment designs can be used for hybrid designs. This design is a combination of an enrichment design where we randomize patients to either the experimental or the control treatment group and a single-arm design in biomarker-negative patients.

**Statistical considerations:** The strength of the hybrid design is that apart from the evaluation of the predictive ability of a biomarker, the feasibility of a prognostic biomarker can also be tested. It can be considered as an advantageous design of the enrichment designs when there is prior evidence showing not only that the control treatment works well for the biomarker-negative population but also a detrimental effect of the experimental treatment for that subgroup or possible excessive toxicity as we do not exclude these patients from the trial as it happens in the enrichment designs.

#### 3.2.4. Biomarker-Strategy Designs

---

Generally, with biomarker-strategy designs, the study population is randomized to treatment strategies as opposed to treatments per se. More precisely, patients are randomized to either a biomarker-based treatment strategy arm where the biomarker is used in deciding on approach to treatment, or to an arm that does not use the biomarker to guide treatment. Consequently, biomarker-strategy designs make a comparison between two strategies—one which uses biomarker information to inform treatment approach and the other that does not.

These designs are also known as biomarker-based strategy designs or signature-based strategy designs and they are composed of four subtypes; (i) biomarker-strategy designs with biomarker assessment in the control arm; (ii) biomarker-strategy designs without biomarker assessment in the control arm; (iii) biomarker-strategy designs with treatment randomization in the control arm and (iv) reverse marker-based strategy designs. Whilst patients randomized to the non-biomarker based strategy arm in the first two design subtypes are allocated the control treatment, in the third design subtype those patients undergo secondary

randomization to either the control or experimental treatment. The fourth design subtype differs from the three aforementioned subtype designs as the non-biomarker based strategy arm is replaced by the reverse marker-strategy arm. The first and second types are similar with the difference being only in terms of ethical/feasibility issues regarding the acquisition of biomarker status at the beginning of the trial.

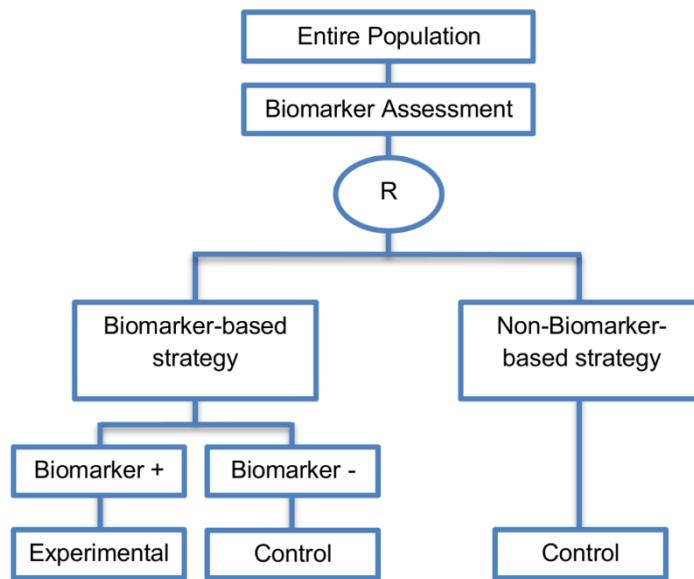
This approach is preferred when the study is planned for a confirmatory phase of a certain biomarker-based strategy allowing for comparison between the biomarker-based strategy and non-biomarker-based strategy.

#### *3.2.4.1. Biomarker-Strategy Design with Biomarker Assessment in the Control Arm*

---

This approach is described in 21 (21%) papers of our review.

**Design:** First, the study population enrolled in the trial is tested for its marker status. Next, patients irrespective of their biomarker status are randomized either to the biomarker-based strategy arm (also referred to as personalized arm) or to the non-biomarker-based strategy arm. In the biomarker-based strategy arm, biomarker-positive patients receive the experimental treatment, whereas, biomarker-negative patients receive the control treatment. Patients who are randomized to the non-biomarker-based strategy arm receive the control treatment irrespective of their biomarker status. A graphical illustration of this design is given in Figure 3.12. This biomarker-strategy design can be extended to more than one experimental treatment. More precisely, this extension is referred to as Individual profile design in literature and was identified in two papers [36, 72] (2%) of our review. This design includes different individual status, e.g., instead of biomarker-positive and biomarker-negative subgroups we can have patients who are positive for biomarker 1, biomarker 2, biomarker n, leading to the selection of personalized treatments, (patients who are positive for biomarker 1 are treated with the corresponding experimental treatment 1, etc.).



**Figure 3.12.** Biomarker-strategy design with biomarker assessment in the control arm. “R” refers to randomization of patients.

**Utility:** This approach is useful when we want to test the hypothesis that the treatment effect based on the biomarker-based strategy approach is superior to that of the standard of care.

**Methodology:** The clinical utility of a biomarker can be evaluated by comparing the two strategy groups. The predictive utility of the marker-based treatment strategy could be assessed by comparing the outcome of all patients in the biomarker-based strategy arm to all patients in the non-biomarker-based strategy arm. Patients in the marker-based strategy arm do not need to be limited to two treatments; in principle, a marker-based strategy involving many biomarkers and many possible treatments could be compared to standard of care treatment.

According to Freidlin et al., 2010 [61], assuming a survival outcome, the required sample size in terms of number of events for this type of biomarker-strategy design in order to reach power  $(1 - \beta)$  at significance level  $\alpha$  (type I error) can be given by

$$D_{strategy I} = 4 \left[ \frac{(z_{a/2} + z_{\beta})}{k \log \theta_1} \right]^2, \quad (3.32)$$

where  $k$  denotes the prevalence of biomarker-positive patients,  $\theta < 1$  denotes the assumed hazard ratio in the biomarker-positive subpopulation and  $z_{a/2}$ ,  $z_{\beta}$  denote the upper  $a/2$ - and upper  $\beta$ -points respectively of a standard normal distribution where  $a$  and  $\beta$  denote the assumed type I error and type II error respectively. According to Freidlin et al. 2010 [61], it is assumed that there is no treatment effect in the biomarker-negative subpopulation (corresponding to a hazard ratio of experimental treatment versus control treatment of 1) and that there is no prognostic effect of the biomarker under the control treatment. Consequently, the overall hazard ratio between experimental and control arms in biomarker-positive patients and biomarker-negative patients can be approximated by  $\exp[k \log \theta + (1 - k) \log 1] = \theta^k$  [61] and this is the reason why the formula which gives the required total number of events ( $D_{strategy}$ ) contains only the hazard ratio of biomarker-positive patients. Freidlin et al., 2010 [61] provided the aforementioned formula assuming that all random assignments use 1:1 randomization.

Additionally, Young et al., 2010 [26] determined the total sample size needed for this type of biomarker-strategy designs when using continuous clinical endpoints by

$$N_{strategy I} = \frac{2(z_{1-a/2} + z_{1-\beta})^2 (\tau_m^2 + \tau_n^2)}{(v_m - v_n)^2}, \quad (3.33)$$

where  $z_{1-a/2}$ ,  $z_{1-\beta}$  denote the lower  $1 - a/2$ - and lower  $1 - \beta$ -points respectively of a standard normal distribution,  $a$  and  $\beta$  denote the assumed type I error and type II error respectively,  $v_m$  and  $v_n$  denote the mean response from the biomarker-based strategy arm and the non-biomarker-based strategy arm respectively, and  $\tau_m^2$ ,  $\tau_n^2$  denote the variance of response for the biomarker-based strategy arm and non-biomarker-based strategy arm respectively. Young et al., 2010 [26] also provided formulae for the aforementioned variances which depend on sensitivity and



specificity of the assay, such that any error in the evaluation of biomarker in the biomarker-based strategy can be accounted for.

For the case of binary outcomes, Eng, 2014 [92] provided the formula for the required sample size for each arm in a test of proportions between the two randomization arms (biomarker-based strategy arm and non-biomarker-based strategy arm). This formula can be given by

$$N_{strategy\ I/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_2(1 - g_2)]}{\Delta_2^2} \quad (3.34)$$

where  $a$  corresponds to the target level,  $1 - \beta$  corresponds to the power,  $g_1$  is the expected response rate in the biomarker-based strategy arm,  $g_2$  is the expected response rate in the non-biomarker-based strategy arm and  $\Delta_2 = g_1 - g_2$ . The expected response rates  $g_1, g_2$  can be found by calculating the formulae  $kr_{A+} + (1 - k)r_{B-}$  and  $r_B$  respectively, the prevalence of biomarker-positive patients corresponds to  $k$  and  $r_{A+}, r_{B-}$  are the assumed response rates of biomarker-positive patients receiving the experimental treatment and biomarker-negative patients receiving the control treatment,  $r_B$  denotes the marginal effect of treatment B (control treatment).

**Statistical considerations:** This type of design is able to inform researchers whether the biomarker is prognostic, since both biomarker positive and negative patients are exposed to the control treatment, but it cannot answer the question of whether the biomarker is predictive since only biomarker positive patients are exposed to the experimental treatment. Additionally, these designs have been criticized by many authors as less efficient than the marker-stratified designs since it is possible for some patients in both the biomarker-based strategy arm and non-biomarker-based strategy arm to be assigned to the same treatment (due to the existence of biomarker-negative patients in both strategy arms the treatment effect can be diluted) and they require a large sample size to detect an overall difference in outcomes between arms. Furthermore, these designs cannot compare experimental treatment to control treatment directly as they are designed to compare not the

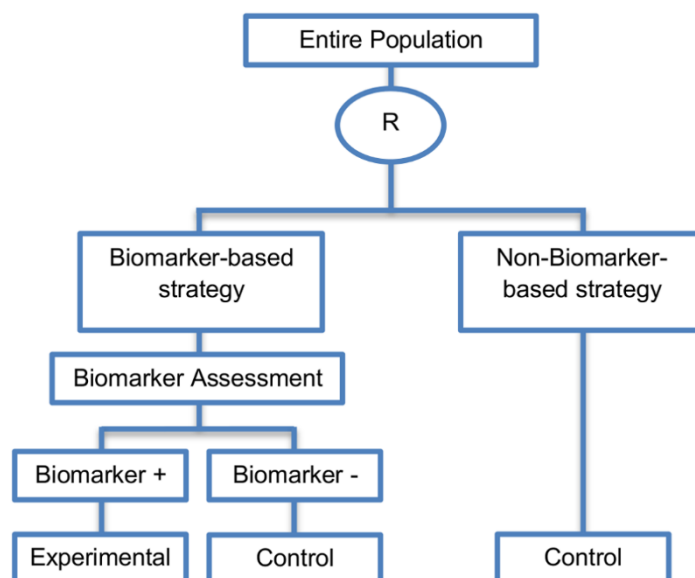
treatments but the biomarker-strategies. Another limitation of these designs is the uncertainty about whether the results which indicate efficacy of the biomarker-directed approach to treatment are caused due to a true effect of the biomarker or due to a treatment effect irrespective of the biomarker status.

#### 3.2.4.2. Biomarker-Strategy Design without Biomarker Assessment in the Control Arm

---

This strategy was identified in 14 papers (14%) of our review.

**Design:** In this approach, patients are again randomized between testing strategies (i.e. biomarker-based strategy and non-biomarker-based strategy) but it differs in terms of the timing of biomarker evaluation. More precisely, first, patients are randomized to either the biomarker-based strategy or to the non-biomarker-based strategy. Next, this design evaluates the biomarkers only in patients who are assigned to the biomarker-based strategy. Patients who are found to be biomarker-positive will receive the experimental treatment and patients who are biomarker-negative will receive the control treatment. On the other hand, the population which is randomized to the non-biomarker-based strategy will receive the control treatment. A graphical illustration of this design is given in Figure 3.13.



**Figure 3.13.** Biomarker-strategy design without biomarker assessment in the control arm. “R” refers to randomization of patients.

**Utility:** This design is useful in situations where it is either not feasible or ethical to test the biomarker in the entire population due to several logistical (e.g., specimens not submitted), technical (e.g., assay failure) or clinical reasons (e.g., tumor inaccessible); thus the biomarker status is obtained only in patients who are tailored to the biomarker-based strategy arm.

**Methodology:** The same mathematical formula for sample size calculation assuming a continuous clinical outcome proposed by Young et al. (2010) [26] and the formula assuming binary outcome proposed by Eng, 2014 [92] for the biomarker-strategy design with biomarker assessment in the control arm could be applied. Further, in terms of survival outcome, the same formula provided for the required number of events in the first version of biomarker-strategy designs (i.e. biomarker-strategy design with biomarker assessment in the control arm) could be considered.

**Statistical considerations:** These designs have the same advantages and limitations as the previously discussed biomarker-strategy design with biomarker assessment in the control arm, e.g., they have been criticized for their lack of efficiency due to the fact that biomarker negative patients are exposed to the control treatment in both arms of the trial. An additional limitation is that the biomarker-positive and biomarker-negative subpopulations might be more imbalanced as compared with the first type of biomarker-strategy design due to the fact that the randomization is performed before the evaluation of biomarker (balancing the randomization is useful to ensure that all randomized patients have tissue available).

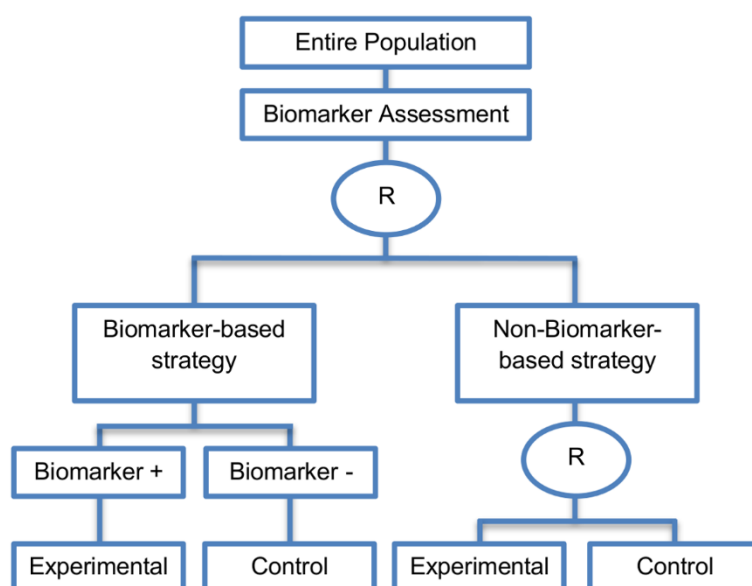
#### *3.2.4.3. Biomarker-Strategy Design with Treatment Randomization in the Control Arm*

---

Sargent and Allegra [108] proposed another version of Biomarker-strategy designs where there is a second randomization between experimental and control treatment in the non-biomarker guided strategy arm. This strategy is referred to in 17 papers (17%) of our review.

**Design:** A graphical illustration of this approach is given in Figure 3.14. The two previously described biomarker-strategy designs can answer the question about

whether the biomarker-based strategy is more effective than standard treatment, irrespective of the biomarker status of the study population, whereas the biomarker-strategy design with treatment randomization in the control treatment is able to inform us about whether the biomarker-based strategy is better than not only the standard treatment but also better than the experimental treatment in the overall population. This is achieved by using a second randomization the ratio of which should be informed by the prevalence of the biomarker in question in the population as a whole to ensure balance between the study arms. Patients are first randomly assigned to either the biomarker-based strategy arm or to the non-biomarker-based strategy arm. Next, patients who are allocated to the non-biomarker-based strategy are again randomized either to the experimental treatment arm or to the standard treatment arm irrespective of their biomarker status. Patients who are allocated to the biomarker-based strategy and who are biomarker-positive are given the experimental treatment and patients who are biomarker-negative are given the control treatment. The clinical utility of the biomarker is evaluated by comparing treatment effect between the biomarker-based strategy arm and non-biomarker-based strategy arm. Such an approach can also identify whether a novel treatment is more effective in the entire population or in a biomarker-defined subgroup only, since both biomarker subgroups are exposed to both treatments.



**Figure 3.14.** Biomarker-strategy design with treatment randomization in the control arm. “R” refers to randomization of patients.

**Utility:** These designs are preferable as compared to the two previously discussed biomarker-strategy designs in cases where there is interest in whether the biomarker is not only prognostic but also predictive.

**Methodology:** Mandrekar and Sargent, 2009 [31] calculated the total required sample size in terms of number of events for the comparison of a survival outcome in the biomarker-based strategy versus the non-biomarker-based strategy. According to them, the required total number of events when using 1:1 randomization to treatment arms is given by

$$D_{strategy\ III} = \frac{4(z_{a/2} + z_{\beta})^2}{\left\{ \log \left[ \frac{2\kappa m_{B+} + 2(1-\kappa)m_{A-}}{\kappa(m_{A+} + m_{B+}) + (1-\kappa)(m_{A-} + m_{B-})} \right] \right\}^2}, \quad (3.35)$$

where  $\kappa$  denotes the prevalence of the biomarker-positive patients,  $m_{A+}, m_{A-}, m_{B+}, m_{B-}$ , denote the median survival for biomarker-positive and biomarker-negative patients receiving control and experimental treatments respectively. Also, the constants  $z_{a/2}, z_{\beta}$  denote the upper  $a/2$ - and upper  $\beta$ -points respectively of a standard normal distribution where  $a$  and  $\beta$  denote the assumed type I error and type II error respectively.

Additionally, Young et al., 2010 [26], considering continuous clinical outcomes, calculated the total sample size by

$$N_{strategy\ III} = \frac{2(z_{1-a/2} + z_{1-\beta})^2 (\tau_m^2 + \tau_{nr}^2)}{(v_m - v_{nr})^2}, \quad (3.36)$$

where  $z_{1-a/2}, z_{1-\beta}$  denote the lower  $1 - a/2$ - and lower  $1 - \beta$ -points respectively of a standard normal distribution,  $a$  and  $\beta$  denote the assumed type I error and type II error respectively,  $v_m$  and  $v_{nr}$  denote the mean response from the biomarker-based strategy arm and the non-biomarker-based strategy arm, and  $\tau_m^2, \tau_{nr}^2$  denote the variance of response for the biomarker-based strategy arm and non-biomarker-based

strategy arm respectively. The only differences in the mathematical formula for the total sample size between this type of biomarker-strategy design and the first and second types mentioned above are the values of  $v_{nr}$  and  $\tau_{nr}^2$ , to reflect the fact that in the non-biomarker-based strategy arm patients are randomly assigned to either the experimental or control treatment. Again, the formulae can be adjusted to account for uncertainty in biomarker assessment.

For the case of binary outcomes, Eng, 2014 [92] provided the formula for the required sample size for each arm in a test of proportions between the two randomization arms (biomarker-based strategy arm and non-biomarker-based strategy arm). This formula can be given by

$$N_{strategy\ III/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_3(1 - g_3)]}{\Delta_3^2} \quad (3.37)$$

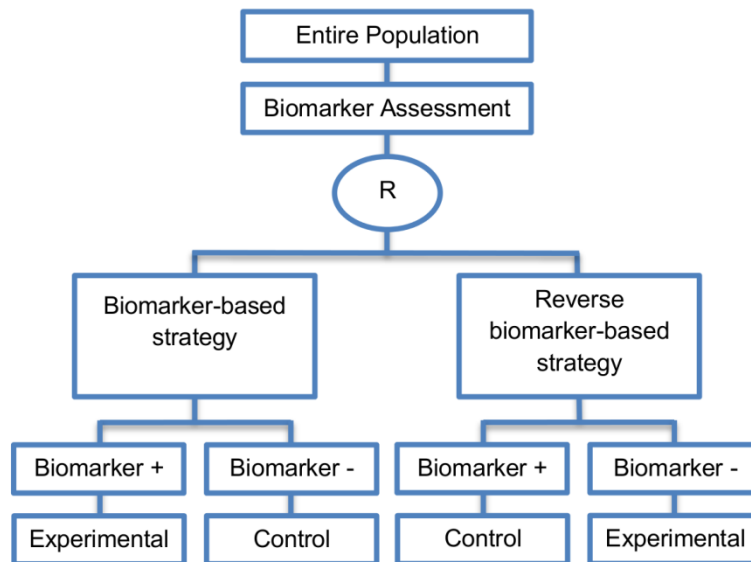
where  $a$  correspond to the target level,  $1 - \beta$  corresponds to the power,  $g_1$  is the expected response rate in the biomarker-based strategy arm,  $g_3$  is the expected response rate in the non biomarker-based strategy arm and  $\Delta_3 = g_1 - g_3$ . The expected response rates  $g_1, g_3$  can be found by calculating the formulae  $kr_{A+} + (1 - k)r_{B-}$  and  $r_A/2 + r_B/2$  respectively,  $r_A$  and  $r_B$  denote the marginal effect of treatment A (experimental treatment) and treatment B (control treatment) respectively.  $r_{A+}, r_{B-}$  are the assumed response rates of biomarker-positive patients receiving the experimental treatment and biomarker-negative patients receiving the control treatment. The prevalence of biomarker-positive patients corresponds to  $k$ .

**Statistical considerations:** Similar to both aforementioned biomarker-strategy designs, the biomarker-strategy design with treatment randomization in the control arm will need larger sample size as compared to the marker-stratified designs. However, one strength is that they allow clarification of whether the results which indicate efficacy of the biomarker-directed approach to treatment are caused due to a true effect of the biomarker or due to a treatment effect irrespective of the biomarker status which does not happen in the first two types of biomarker-strategy designs.

#### 3.2.4.4. Reverse Marker-Based Strategy Design

Eng, 2014 [92] proposed another version of biomarker-strategy designs where the non-biomarker-based strategy arm which is included in the three aforementioned subtypes of biomarker-strategy designs is replaced by the reverse marker-strategy arm. This strategy is referred to in four papers (4%) of our review.

**Design:** A graphical illustration of this approach is given in Figure 3.15. In this design patients are randomized either to the biomarker-based strategy arm or the reverse biomarker-based strategy arm. As in the previous three biomarker-strategy subtype designs, patients who are allocated to the biomarker-strategy arm receive the experimental treatment if they are biomarker-positive whereas biomarker-negative patients receive the control treatment. By contrast, patients who are randomly assigned to the reverse biomarker-based strategy arm receive control treatment if they are biomarker-positive, whereas biomarker-negative patients receive experimental treatment.



**Figure 3.15.** Reverse Marker-Based strategy design. “R” refers to randomization of patients.

**Utility:** Reverse marker-based strategy is a more efficient strategy as compared to the first and third biomarker-strategy subtype design for testing the interaction hypothesis of treatment and biomarker. This design should be used in cases where

prior evidence indicates that both experimental and control treatment are effective in treating patients but the optimal strategy has not yet been identified.

**Methodology:** This subtype design is balanced (i.e. the randomization frequencies for each treatment are equal independent of the prevalence of the biomarker) and it is powered to evaluate the interaction between treatment and biomarker. For the case of binary outcomes, Eng, 2014 [92] provided the formula for the required sample size for each arm in a test of proportions between the two randomization arms (biomarker-based strategy arm and reverse biomarker-based strategy arm). This formula can be given by

$$N_{strategy\ IV/arm} = \frac{(z_a + z_{1-\beta})^2 [g_1(1 - g_1) + g_4(1 - g_4)]}{\Delta_4^2} \quad (3.38)$$

where  $a$  correspond to the target level,  $1 - \beta$  corresponds to the power,  $g_1$  is the expected response rate in the biomarker-based strategy arm,  $g_4$  is the expected response rate in the reverse biomarker-based strategy arm and  $\Delta_4 = g_1 - g_4$ . The expected response rates  $g_1, g_4$  can be found by calculating the formulae  $kr_{A+} + (1 - k)r_{B-}$  and  $kr_{B+} + (1 - k)r_{A-}$  respectively,  $r_{A+}, r_{B-}$  are the assumed response rates of biomarker-positive patients receiving the experimental treatment and biomarker-negative patients receiving the control treatment and  $r_{A-}, r_{B+}$  are the assumed response rates of biomarker-negative patients receiving the experimental treatment and biomarker-positive patients receiving the control treatment. The prevalence of biomarker-positive patients corresponds to  $k$ .

**Statistical considerations:** This design enables the evaluation of the interaction between the biomarker and different treatments and can estimate directly the marker-strategy response rate. Additionally, this subtype design allows the estimation of the effect size of the experimental treatment compared to the control treatment for each biomarker-defined subgroup separately. Also, there is no chance that the same treatment will be tailored to biomarker-positive patients who are randomized either to the biomarker-based strategy arm or the reverse marker strategy (i.e. biomarker-positive patients in the biomarker-based strategy will be

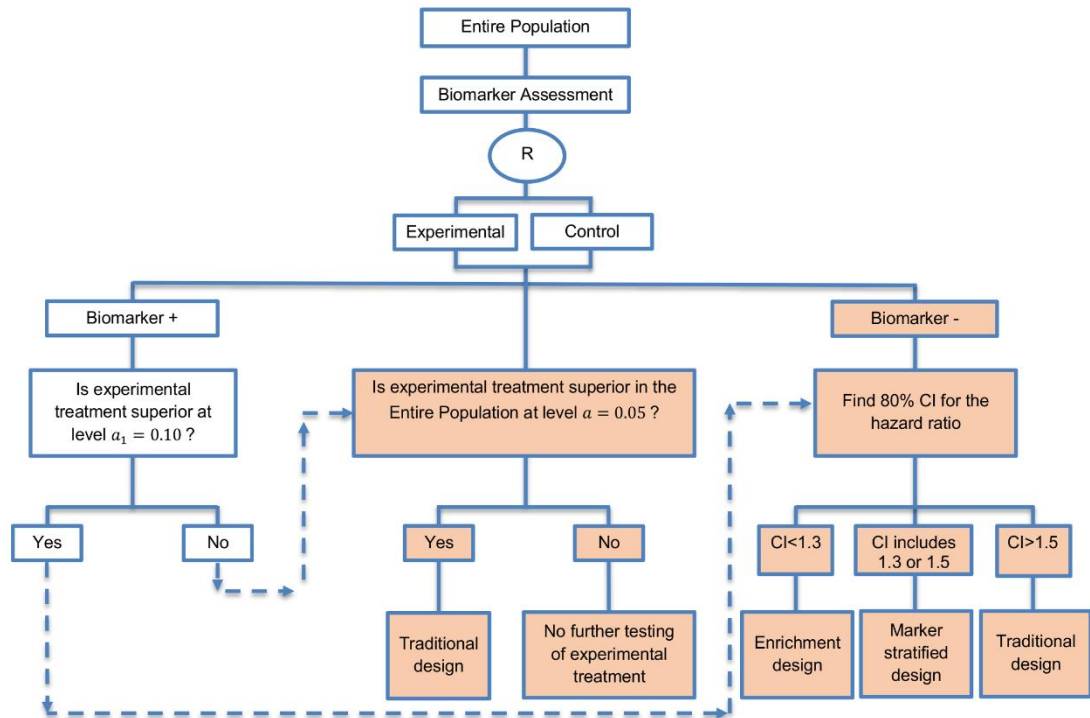


given only the experimental treatment and biomarker-positive patients in the reverse marker strategy arm will be given only the control treatment). Also, there is no possibility of the same treatment assignment to biomarker-negative patients who are randomly assigned to the two biomarker-based strategy arms (i.e. biomarker-negative patients in the marker-based strategy arm will be treated with the control treatment, whereas biomarker-negative patients in the reverse marker strategy arm will be treated with the experimental treatment). According to Eng, 2014 [92] who compared the reverse marker-based strategy design with the first (i.e. biomarker-strategy design with biomarker assessment in the control arm) and third (i.e. biomarker-strategy design with treatment randomization in the control arm) subtype of biomarker-strategy designs in the case of binary outcomes, the effect size in order to make a comparison of the different treatment strategy arms would be larger than in the first and third subtype designs. Furthermore, it has been shown by Eng, 2014 that in situations where a randomly chosen treatment has a better than 7% response rate, the reverse marker-based strategy design works better as compared to the third biomarker-strategy subtype (i.e. Biomarker-strategy design with treatment randomization in the control arm). It has also been demonstrated that this novel design is more than four times more efficient in order to test the interaction between treatment and biomarker compared to Biomarker-strategy design with biomarker assessment in the control arm, Biomarker-strategy design with randomization in the control arm and the marker stratified design. Eng, 2014 demonstrated the benefits of the Reverse Marker-Based strategy design with the aim to assess the interaction between treatment and biomarker. However, Baker, 2014 [93] stated that other designs than the Reverse Marker-Based strategy design would be more appropriate in order to investigate questions which include treatment effect of biomarker-defined subgroups and biomarker-based strategy arms.

### 3.2.5. Other Designs

#### 3.2.5.1. A Randomized Phase II Trial Design with Biomarker Proposed by Freidlin et al., 2012

Freidlin et al., 2012 [71] proposed a biomarker-guided Phase II clinical trial design in which it, when completed, recommends which type of Phase III trial should be used. These recommendations for a Phase III trial are the following: (i) enrichment design; (ii) marker-stratified design; (iii) a traditional trial design without a biomarker; or (iv) drop consideration of the experimental treatment. A graphical illustration of this design is given in Figure 3.16.



**Figure 3.16.** Randomized Phase II trial design with biomarkers. “R” refers to randomization of patients. CI refers to the confidence interval. Uncolored boxes are referred to the first stage of the trial and colored boxes are referred to the second stage of the trial. Different stages refer to the analysis and not to the trial design.

**Design:** For this type of randomized Phase II trial, it is assumed that the experimental treatment will be more beneficial among biomarker-positive patients than biomarker-negative patients without ruling out the efficacy of the novel

treatment in biomarker-negative patients. The intermediate endpoint of progression-free survival (PFS) is used which is able not only to give the results earlier but also to target larger treatment effects as compared to overall survival (OS) endpoint.

The design starts by comparing the experimental treatment with the control treatment in the biomarker-positive subgroup using a one-sided level of significance  $\alpha_1 = 0.10$ . The null hypothesis is that the progression-free survival for biomarker-positive patients is the same for both experimental and control treatment arm ( $HR_{0,biom+} \leq 1$  vs.  $HR_{1,biom+} > 1$ ). Next, if the null hypothesis is rejected, which means that the experimental treatment is better than the control treatment in the biomarker-positive subgroup we continue with the calculation of an 80% two-sided confidence interval (CI) for the hazard ratio (control vs experimental) in the biomarker-negative subpopulation. Three decisions are made according to the values of the CI: (i) if the entire CI is less than 1.3 then we can continue with a Phase III enrichment design; (ii) if the CI includes the values 1.3 or 1.5 then we can continue with a Phase III marker-stratified design and (iii) if the entire CI is greater than 1.5 then it seems that the biomarker is not useful as the novel treatment benefits both the biomarker-negative and biomarker-positive patients, thus, the biomarker should be dropped and a traditional randomized Phase III design should be conducted. Otherwise, if the null hypothesis is not rejected at the one-sided significance  $\alpha_1 = 0.10$  (meaning that the experimental treatment is not better than the experimental treatment in the biomarker-positive subgroup), then we continue with the comparison of treatments in the overall study population at one-sided level of significance  $\alpha = 0.05$ . If the null hypothesis of no treatment effect in the entire population is rejected, then the authors recommend to drop the biomarker and to continue with a traditional randomized Phase III trial due to the fact that the biomarker seems to be useless. On the other hand, if the null hypothesis is not rejected, the experimental treatment should not be tested further as it does not seem to be effective.

**Utility:** This design should be used when we want to conduct a Phase II randomized trial which allows decisions to be made about which type of Phase III

biomarker-guided trial to proceed with. It is appropriate when there is prior evidence that the novel treatment benefits mostly the biomarker-positive patients without ruling out treatment effect in biomarker-negative patients.

**Methodology:** Freidlin et al., 2012 [71] have provided an online tool for calculating the sample size which can be found on the following website <http://brb.nci.nih.gov/Data/FreidlinB/RP2BM> [115]. In order for a sample size to be estimated, the following information is required: (i) the significance levels for testing the treatment effect in the biomarker-positive subgroup and in the entire population; (ii) cut-offs and confidence intervals for the hazard ratio in the biomarker-negative subgroup; (iii) the prevalence of biomarker-positive patients; (iv) the median progression-free survival in each treatment arm in each biomarker-defined subgroup and (v) the accrual parameters. Regarding the accrual parameters, the author specifies the minimum sample size for biomarker-positive patients for which the accrual continues until this number is reached, the maximum number of over-accrual in biomarker-positive subgroup for which the accrual to the entire population stops after this number is reached and the maximum accrual number in biomarker-negative patients for which the accrual to this biomarker-defined subgroup stops when this number is reached.

**Statistical considerations:** In real life, it might not be possible to obtain the biomarker status for the entire population. If the biomarker status is unknown for some patients, then these individuals could be included in the analysis of the overall population. More precisely, in case that the proportions of patients with unknown biomarker status is low, the randomization of them to either the experimental or the control treatment could be considered in the second stage of this Phase II trial where we test the treatment effectiveness in the entire population. Another statistical consideration is that researchers should take into account the adjustment for inflation in Phase III type I error as the chosen Phase III trial design depends on the performance of the aforementioned randomized Phase II trial. Additionally, the authors suggest generally that in cases where it seems that the control treatment has been shown more beneficial, an aggressive interim inefficacy/futility should be used,

i.e. when the estimated hazard ratio of control treatment versus the experimental treatment is equal or less than one when half of the required number of events have been observed, then the accrual should stop to that biomarker-defined subgroup.

### 3.3. Discussion

---

A number of biomarker-guided trial designs have been proposed in the past decade, including both biomarker-guided adaptive and non-adaptive trial designs. We have undertaken a comprehensive review of the literature using an in-depth search strategy to report on the biomarker-guided designs proposed to date, with a view to providing the research community with clarity in definition, methodology and terminology of the various trial designs. The review is split in two parts due to its size; the first part of the review is focused on adaptive designs which are extensively discussed in Chapter 2 [35], whereas, herein we focus on non-adaptive designs which incorporate biomarkers.

The review has demonstrated ambiguity and confusion regarding the biomarker-guided non-adaptive designs proposed by different authors. For example, different authors described the same trial design but with different names (see Table 3.1). In this review, we focus on 5 main types of such designs including their subtypes and variations. Knowledge on how to implement and analyse these designs are essential in testing the effectiveness of a biomarker-guided approach to treatment; hence, a comprehensive review giving this knowledge is essential for the research community. In our in-depth study, we provide researchers with analytical information of these study designs not only in terms of their utility, advantages and limitations but also in terms of their methodology. In addition, a graphical illustration for each biomarker-guided design is given. A guidance document by Tajik et al., 2012 [116] regarding the evaluation of putative biomarkers in randomized clinical trials came to our knowledge by personal communication as we were not able to identify it during our literature search.

The non-adaptive designs do not allow modifications of important aspects of the trial such as refinement of the existing study population, treatment assignment, study endpoints, study duration, etc. In non-adaptive designs, all these factors are defined before the initiation of the study and they are kept fixed during the course of the clinical trial. However, there is a great potential of failure when implementing such conventional designs due to potential wrong design assumptions of the key aspects of the study that might be made before the conduct of the trial. Hence, an adaptive design clinical study which allows on-going adaptations based on accumulating study data from interim analysis might hold advantageous position as compared to the non-adaptive trial design due to its flexibility. However, before implementing an adaptive design a lot of issues should be taken into careful consideration by research teams in order to prove that there are good reasons for conducting such designs. Regulatory and logistical issues, requirement of additional efforts for the achievement of the design, potential difficulties, possible increased cost and time, statistical challenges including the potential increase of the chance of a false conclusion that the treatment is effective (inflation of Type I error) and whether the adaptation process has led to positive study results that are difficult to interpret irrespective of having control of Type I error should be considered [129]. A recent paper by Dimairo et al., 2015 [130] refers to a number of obstacles and barriers when implementing adaptive designs in practice. Several key stakeholders in clinical trials research have been interviewed (i.e. UK Clinical Trials Units directors, funding board and panel members, statisticians, regulators, chief investigators, data monitoring committee members and health economists) expressing difficulties of adaptive designs. Lack of appropriate knowledge and familiarity of these designs in the scientific community, insufficient time and funding structure, additional work required due to the complexity of such designs and the needed statistical expertise and appropriate software are some of the highlighted difficulties mentioned in the paper of Dimairo et al., 2015 [130]. In addition, this study includes the characterisation of potential benefits of an adaptive design to patients, clinical trials as well as funders.

The different designs proposed so far for biomarker-guided designs, both non-adaptive designs which remain an appealing approach to a great extent mainly due to their simplicity and adaptive designs which are more flexible, need to be further explored by the research community, as the proper choice and use of such designs can result in a great increase in the efficiency of a trial and expedite the development of novel treatments.

The characteristics and methodology of the five main designs and their subtypes are discussed in the current chapter, whilst information on their variations are summarized in Appendix B. Additional references for these variations and the literature review search strategy are provided in [131, 132].

In this chapter we presented and discussed in detail the different types of non-adaptive trial designs. To ensure that the guidance and graphical representations of Chapter 2 and 3 are easily accessible to stakeholders, an interactive web resource to host this information was also developed and is presented in Chapter 4.

### 3.4. References

---

1. George SL. Statistical issues in translational cancer research. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; (19):5954-8. doi: 10.1158/1078-0432.CCR-07-4537.
2. Chabner B. Advances and challenges in the use of biomarkers in clinical trials. *Clinical advances in hematology & oncology*. 2008; 6(1):42-3.
3. Group BDW. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*. 2001; 69(3):89-95. doi: 10.1067/mcp.2001.113989.
4. Shi Q, Mandrekar SJ, Sargent DJ. Predictive biomarkers in colorectal cancer: usage, validation, and design in clinical trials. *Scandinavian journal of gastroenterology*. 2012; 47(3):356-62. doi: 10.3109/00365521.2012.640836.

5. Pihlstrom BL, Barnett ML. Design, operation, and interpretation of clinical trials. *Journal of dental research*. 2010; 89(8):759-72. doi: 10.1177/0022034510374737.
6. Rigatto C, Barrett BJ. Biomarkers and surrogates in clinical studies. *Methods in molecular biology* (Clifton, NJ). 2009; 473:137-54. doi: 10.1007/978-1-59745-385-1\_8.
7. Mandrekar SJ, An M-W, Sargent DJ. A review of phase II trial designs for initial marker validation. *Contemporary clinical trials*. 2013; 36(2):597-604. doi: 10.1016/j.cct.2013.05.001.
8. Karuri SW, Simon R. A two-stage Bayesian design for co-development of new drugs and companion diagnostics. *Statistics in medicine*. 2012; 31(10):901-14. doi: 10.1002/sim.4462.
9. Matsui S. Genomic biomarkers for personalized medicine: development and validation in clinical studies. *Computational and mathematical methods in medicine*. 2013; 2013:865980. doi: 10.1155/2013/865980.
10. Buyse M, Michiels S. Omics-based clinical trial designs. *Current opinion in oncology*. 2013; 25(3):289-95. doi: 10.1097/CCO.0b013e32835ff2fe.
11. Wu W, Shi Q, Sargent DJ. Statistical considerations for the next generation of clinical trials. *Seminars in oncology*. 2011; 38(4):598-604. doi: 10.1053/j.seminoncol.2011.05.014.
12. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2005;23(9):2020-7. doi: 10.1200/JCO.2005.01.112.
13. Chen JJ, Lu T-P, Chen D-T, Wang S-J. Biomarker adaptive designs in clinical trials. *Translational Cancer Research*. 2014; 3(3):279-92.



14. Freidlin B, Sun Z, Gray R, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2013; 31(25):3158-61. doi: 10.1200/JCO.2012.48.3826.
15. Goshio M, Nagashima K, Sato Y. Study designs and statistical analyses for biomarker research. *Sensors (Basel, Switzerland)*. 2012; (7):8966-86. doi: 10.3390/s120708966.
16. Ming-Wen An SJM, Daniel JS. Biomarkers-guided targeted drugs: new clinical trials design and practice necessity. *Advances in Personalized Cancer Management*. 2011:30-41. doi: 10.2217/ebo.11.87.
17. Buyse M. Towards validation of statistically reliable biomarkers. *European Journal of Cancer Supplements*. 2007; 5(5):89-95. doi: 10.1016/S1359-6349(07)70028-9.
18. Lee CK, Lord SJ, Coates AS, Simes RJ. Molecular biomarkers to individualise treatment: assessing the evidence. *The Medical journal of Australia*. 2009; 190(11):631-6.
19. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized medicine*. 2010; 7(1):33-47. doi: 10.2217/pme.09.49.
20. Fraser GAM, Meyer RM. Biomarkers and the design of clinical trials in cancer. *Biomarkers in medicine*. 2007; 1(3):387-97. doi: 10.2217/17520363.1.3.387.
21. Mandrekar SJ, Sargent DJ. Design of clinical trials for biomarker research in oncology. *Clinical investigation*. 2011; 1(12):1629-36. doi: 10.4155/CLI.11.152.
22. Simon R. Advances in clinical trial designs for predictive biomarker discovery and validation. *Current Breast Cancer Reports*. 2009; 1(4):216-21. doi: 10.1007/s12609-009-0030-4.

23. Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*. 2013; 105(22):1677-83. doi: 10.1093/jnci/djt282.
24. Bradley E. Incorporating biomarkers into clinical trial designs: points to consider. *Nature biotechnology*. 2012; 30(7):596-9. doi: 10.1038/nbt.2296.
25. Beckman RA, Clark J, Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature reviews Drug discovery*. 2011; (10):735-48. doi: 10.1038/nrd3550.
26. Young KY, Laird A, Zhou XH. The efficiency of clinical trial designs for predictive biomarker validation. *Clinical trials (London, England)*. 2010; 7(5):557-66. doi: 10.1177/1740774510370497.
27. Lee JJ, Xuemin G, Suyu L. Bayesian adaptive randomization designs for targeted agent development. *Clinical trials (London, England)*. 2010; 7(5):584-96. doi: 10.1177/1740774510373120.
28. Simon R. Clinical trials for predictive medicine: new challenges and paradigms. *Clinical trials (London, England)*. 2010; 7(5):516-24. doi: 10.1177/1740774510366454.
29. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points--the challenge of statistical validation. *Nature reviews Clinical oncology*. 2010; 7(6):309-17. doi: 10.1038/nrclinonc.2010.43.
30. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2009; 27(24):4027-34. doi: 10.1200/JCO.2009.22.3701.
31. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal of biopharmaceutical statistics*. 2009; 19(3):530-42. doi: 10.1080/10543400902802458.

32. Hoering A, Leblanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(14):4358-67. doi: 10.1158/1078-0432.CCR-08-0288.
33. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. *Nature reviews Drug discovery*. 2012; 11(3):201-14. doi: 10.1038/nrd3651.
34. Chow S-C. Adaptive clinical trial design. *Annual review of medicine*. 2014; 65:405-15. doi: 10.1146/annurev-med-092012-112310.
35. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *PLoS ONE*. 2016; 11(2):e0149803. doi: 10.1371/journal.pone.0149803.
36. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2013; 19(17):4578-88. doi: 10.1158/1078-0432.CCR-12-3722.
37. Lader EW, Cannon CP, Ohman EM, Newby LK, Sulmasy DP, Barst RJ, et al. The clinician as investigator: participating in clinical trials in the practice setting: Appendix 1: fundamentals of study design. *Circulation*. 2004; 109(21):e302-4.
38. Stingl Kirchheiner JC, Brockmöller J. Why, when, and how should pharmacogenetics be applied in clinical studies?: current and future approaches to study designs. *Clinical pharmacology and therapeutics*. 2011; 89(2):198-209.
39. Sambucini V. A Bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine*. 2008; 27(8):1199-224.
40. Ang M-K, Tan S-B, Lim W-T. Phase II clinical trials in oncology: are we hitting the target? *Expert review of anticancer therapy*. 2010; 10(3):427-38. doi: 10.1586/era.09.178.

41. Farley J, Rose PG, Farley J, Rose PG. Trial design for evaluation of novel targeted therapies. *Gynecologic Oncology*. 2010; 116(2):173.
42. Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, et al. Evaluating many treatments and biomarkers in oncology: a new design. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2013; 31(36):4562-8. doi: 10.1200/JCO.2013.50.7905.
43. Hodgson DR, Wellings R, Harbron C. Practical perspectives of personalized healthcare in oncology. *New Biotechnology*. 2012; 29(6):656-64.
44. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. *Clinical Trials*. 2010; 7(5):567-73.
45. Galanis E, Wu W, Sarkaria J, Chang SM, Colman H, Sargent D, et al. Incorporation of biomarker assessment in novel clinical trial designs: personalizing brain tumor treatments. *Current oncology reports*. 2011; 13(1):42-9. doi: 10.1007/s11912-010-0144-x.
46. Van Schaeybroeck S, Allen WL, Turkington RC, Johnston PG. Implementing prognostic and predictive biomarkers in CRC clinical trials. *Nature reviews Clinical oncology*. 2011; 8(4):222-32. doi: 10.1038/nrclinonc.2011.15.
47. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert review of molecular diagnostics*. 2011; 11(2):171-82. doi: 10.1586/erm.10.120.
48. Sparano JA, Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology: official journal of the American Society of Clinical Oncology*. 2008; 26(5):721-8.
49. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature reviews Clinical oncology*. 2014; 11(2):81-90. doi: 10.1038/nrclinonc.2013.218.

50. Simon R, Polley E. Clinical trials for precision oncology using next-generation sequencing. *Personalized Medicine*. 2013; 10:485-95. doi: 10.2217/pme.13.36.
51. Baker SG, Kramer BS, Sargent DJ, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. *Discovery medicine*. 2012; 13(70):187-92.
52. Buch MH, Pavitt S, Parmar M, Emery P. Creative trial design in RA: optimizing patient outcomes. *Nature Reviews Rheumatology*. 2013; 9(3):183-94.
53. Simon R. Clinical trials for predictive medicine. *Statistics in medicine*. 2012; 31(25):3031-40. doi: 10.1002/sim.5401.
54. Scher HI, Nasso SF, Rubin EH, Simon R. Adaptive clinical trial designs for simultaneous testing of matched diagnostics and therapeutics. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2011; 17(21):6634-40. doi: 10.1158/1078-0432.CCR-11-1105.
55. Sato Y, Laird NM, Yoshida T. Biostatistic tools in pharmacogenomics - advances, challenges, potential. *Current pharmaceutical design*. 2010; 16(20):2232-40.
56. Mandrekar SJ, Sargent DJ, Mandrekar S, Sargent D. Genomic advances and their impact on clinical trial design. *Genome medicine: Medicine in the post-genomic era*. 2009; 1(7):1.
57. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert opinion on medical diagnostics*. 2008; 2(6):721-9. doi: 10.1517/17530059.2.6.721.
58. Dobbin KK. Statistical design and evaluation of biomarker studies. *Methods in molecular biology*. 2014; 1102:667-77.
59. Ananthakrishnan R, Menon S. Design of oncology clinical trials: a review. *Critical reviews in oncology/hematology*. 2013; 88(1):144-53. doi: 10.1016/j.critrevonc.2013.03.007.

60. Simon R. The use of genomics in clinical trial design. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(19):5984-93. doi: 10.1158/1078-0432.CCR-07-4531.
61. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. *JNCI: Journal of the National Cancer Institute*. 2010; 102(3):152-60.
62. Johnson DR, Galanis E. Incorporation of prognostic and predictive factors into glioma clinical trials. *Current oncology reports*. 2013; 15(1):56-63.
63. Sparano J. TAILORx: Trial Assigning Individualized Options for Treatment (Rx). *Clinical Breast Cancer*. 2006; 7(4).
64. Di Maio M, Gallo C, De Maio E, Morabito A, Piccirillo MC, Gridelli C, et al. Methodological aspects of lung cancer clinical trials in the era of targeted agents. *Lung cancer (Amsterdam, Netherlands)*. 2010; 67(2):127-35. doi: 10.1016/j.lungcan.2009.10.001.
65. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005; 24(3):329-39.
66. Collette L, Bogaerts J, Suciú S, Fortpied C, Gorlia T, Coens C, et al. Statistical methodology for personalized medicine: New developments at EORTC Headquarters since the turn of the 21st Century. *European journal of cancer supplements*. 2012; 10(1):13.
67. Mandrekar SJ, Sargent DJ. All-comers versus enrichment design strategy in phase II trials. *Journal of Thoracic Oncology*. 2011; 6(4):658-60.
68. Simon R. Development and validation of biomarker classifiers for treatment selection. *Journal of Statistical Planning and Inference*. 2008; 138:308-20. doi: 10.1016/j.jspi.2007.06.010. PubMed PMID: S037837580700242X.

69. Freidlin B, Korn EL, Gray R. Marker Sequential Test (MaST) design. *Clinical trials* (London, England). 2014; 11(1):19-27. doi: 10.1177/1740774513503739.
70. Wason J, Marshall A, Dunn J, Stein RC, Stallard N. Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *British journal of cancer*. 2014; 110(8):1950-7. doi: 10.1038/bjc.2014.156.
71. Freidlin B, McShane LM, Polley M-YC, Korn EL. Randomized phase II trial designs with biomarkers. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2012; 30(26):3304-9. doi: 10.1200/JCO.2012.43.3946.
72. Ziegler A, Koch A, Krockenberger K, Grosshennig A. Personalized medicine using DNA biomarkers: a review. *Human Genetics*. 2012; 131(10):1627-38.
73. Freidlin B, Korn EL. Biomarker-adaptive clinical trial designs. *Pharmacogenomics*. 2010; 11(12):1679-82. doi: 10.2217/pgs.10.153.
74. Eickhoff JC, Kim K, Beach J, Kolesar JM, Gee JR. A Bayesian adaptive design with biomarkers for targeted therapies. *Clinical trials* (London, England). 2010; 7(5):546-56. doi: 10.1177/1740774510372657.
75. Ferraldeschi R, Attard G, de Bono JS. Novel strategies to test biological hypotheses in early drug development for advanced prostate cancer. *Clinical chemistry*. 2013; 59(1):75-84. doi: 10.1373/clinchem.2012.185157.
76. Coyle VM, Johnston PG. Genomic markers for decision making: what is preventing us from using markers? *Nature reviews Clinical oncology*. 2010; 7(2):90-7. doi: 10.1038/nrclinonc.2009.214.
77. Chen CF, Lin JR, Liu JP. Statistical inference on censored data for targeted clinical trials under enrichment design. *Pharmaceutical Statistics*. 2013; 12(3):165-73.
78. Liu JP, Lin JR. Statistical methods for targeted clinical trials under enrichment design. *Journal of the Formosan Medical Association*. 2008; 107(12 Suppl):35-42.

79. Scheibler F, Zumbé P, Janssen I, Viebahn M, Schröer-Günther M, Grosselfinger R, et al. Randomized controlled trials on PET: a systematic review of topics, design, and quality. *The Journal of Nuclear Medicine*. 2012; 53(7):1016-25.
80. An M-W, Mandrekar SJ, Sargent DJ. A 2-stage phase II design with direct assignment option in stage II for initial marker validation. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2012; 18(16):4225-33. doi: 10.1158/1078-0432.CCR-12-0686.
81. Zheng G, Wu CO, Yang S, Waclawiw MA, DeMets DL, Geller NL. NHLBI clinical trials workshop: an executive summary. *Statistics in Medicine*. 2012; 31(25):2938.
82. Bria E, Di Maio M, Carlini P, Cuppone F, Giannarelli D, Cognetti F, et al. Targeting targeted agents: open issues for clinical trial design. *Journal of Experimental & Clinical Cancer Research*. 2009; 28:66.
83. French B, Joo J, Geller NL, Kimmel SE, Rosenberg Y, Anderson JL, et al. Statistical design of personalized medicine interventions: The Clarification of Optimal Anticoagulation through Genetics (COAG) trial. *Trials*. 2010; 11(1):108.
84. Lin J-A, He P. Reinventing clinical trials: a review of innovative biomarker trial designs in cancer therapies. *British medical bulletin*. 2015; 114(1):17-27. doi: 10.1093/bmb/ldv011.
85. Renfro LA, Mallick H, An M-W, Sargent DJ, Mandrekar SJ. Clinical trial designs incorporating predictive biomarkers. *Cancer Treatment Reviews*. 2016; 43:74-82. doi: <http://dx.doi.org/10.1016/j.ctrv.2015.12.008>.
86. Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics*. 2016.
87. Simon R, Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics Journal*. 2006; 6(3):166-73.



88. European Medicines Agency. Reflection paper on methodological issues associated with pharmacogenomic biomarkers in relation to clinical development and patient selection: London; 2011 [accessed on 10 October 2015]. Available online: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2011/07/WC500108672.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/07/WC500108672.pdf).
89. Lai TL, Liao OY-W, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemporary clinical trials*. 2013; 36(2):651-63. doi: 10.1016/j.cct.2013.08.007.
90. Foley RN. Analysis of randomized controlled clinical trials. *Methods in molecular biology*. 2009; 473:113-26.
91. Tajik P, Bossuyt PM. Genomic markers to tailor treatments: waiting or initiating? *Human Genetics*. 2011; 130(1):15-8.
92. Eng KH. Randomized reverse marker strategy design for prospective biomarker validation. *Statistics in Medicine*. 2014; 33(18):3089-99. doi: 10.1002/sim.6146.
93. Baker SG. Biomarker evaluation in randomized trials: addressing different research questions. *Statistics in Medicine*. 2014; 33(23):4139-40. doi: 10.1002/sim.6202.
94. Matsui S, Choai Y, Nonaka T. Comparison of Statistical Analysis Plans in Randomize-All Phase III Trials with a Predictive Biomarker. *Clinical Cancer Research*. 2014; 20(11):2820-30. doi: 10.1158/1078-0432.CCR-13-2698.
95. Cappuzzo F, Ciuleanu T, Stelmakh L, Cicens S, Szczésna A, Juhász E, et al. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study. *Lancet Oncology (Science Direct)*. 2010; 11(6):521.
96. Hoffmann-La Roche. A Randomized, Double-blind Study to Evaluate the Effect of Tarceva or Placebo Following Platinum-based CT on Overall Survival and Disease Progression in Patients With Advanced, Recurrent or Metastatic NSCLS Who Have

Not Experienced Disease Progression or Unacceptable Toxicity During Chemotherapy: ClinicalTrials.gov; 2007 [accessed on 10 October 2015]. Available online:

<https://clinicaltrials.gov/ct2/show/NCT00556712?term=NCT00556712&rank=1>.

97. Choai Y, Matsui S. Estimation of treatment effects in all-comers randomized clinical trials with a predictive marker: Estimating Treatment Effects in Marker-Based Randomized Trials. *Biometrics*. 2015; 71(1):25.

98. Wang S-J, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical statistics*. 2007; 6(3):227-44. doi: 10.1002/pst.300.

99. Cree IA, Kurbacher CM, Lamont A, Hindley AC, Love S. A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. *Anti-Cancer Drugs*. 2007; 18(9):1093.

100. Cobo M, Isla D, Massuti B, Montes A, Sanchez JM, Provencio M, et al. Customizing Cisplatin Based on Quantitative Excision Repair Cross-Complementing 1 mRNA Expression: A Phase III Trial in Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology: official journal of the American Society of Clinical Oncology*. 2007; 25(19):2747.

101. Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. *Journal of Clinical Epidemiology*. 2009; 62(4):364.

102. Wang S-J. Biomarker as a classifier in pharmacogenomics clinical trials: a tribute to 30th anniversary of PSI. *Pharmaceutical statistics*. 2007; 6(4):283-96. doi: 10.1002/pst.316.

103. Cho D, McDermott D, Atkins M. Designing clinical trials for kidney cancer based on newly developed prognostic and predictive tools. *Current urology reports*. 2006; 7(1):8-15.
104. af Geijerstam JL, Oredsson S, Britton M. Medical outcome after immediate computed tomography or admission for observation in patients with mild head injury: randomised controlled trial. *British Medical Journal*. 2006; 333(7566):465.
105. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology*. 2012; 65(3):282.
106. Mandrekari SJ, Grothey A, Goetz MP, Sargent DJ. Clinical trial designs for prospective validation of biomarkers. *American Journal of Pharmacogenomics*. 2005; 5(5):317-25.
107. Therasse P, Carbonnelle S, Bogaerts J. Clinical trials design and treatment tailoring: general principles applied to breast cancer research. *Critical Reviews in Oncology - Hematology*. 2006; 59(2):98-105.
108. Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Seminars in oncology*. 2002; 29(3):222-30.
109. Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Medical Research Methodology*. 2016; 16(1). doi: 10.1186/s12874-016-0122-6.
110. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases*. 1981; 34(9-10).

111. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials: supplement and correction. *Clinical Cancer Research*. 2006;12: 3229.
112. Simon R, Maitournam A. Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials. *Clinical Cancer Research*. 2004; 10(20):6759.
113. Biomarker Targeted Randomized Design [accessed on 15 September 2016]. Available online: <http://brb.nci.nih.gov/brb/samplesize/td.html>.
114. Harrington RA. Designs for clinical trials: perspectives on current issues. New York, NY: Springer Science+Business Media, LLC; 2012.
115. Freidlin B. Randomized phase II trial designs with biomarkers [accessed on 15 September 2016]. Available online: <http://brb.nci.nih.gov/Data/FreidlinB/RP2BM>.
116. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Evaluating Putative Predictive Biomarkers in Randomized Clinical Trials 2012 [accessed on 15 September 2016]. Available online: [http://www.zonmw.nl/fileadmin/documenten/DO\\_Farmacotherapie\\_Dure\\_Weesgeenesmiddelen/HTA\\_pharmacotherapy\\_predictive\\_markers\\_guidance\\_document.pdf](http://www.zonmw.nl/fileadmin/documenten/DO_Farmacotherapie_Dure_Weesgeenesmiddelen/HTA_pharmacotherapy_predictive_markers_guidance_document.pdf).
117. Zaslavsky BG, Scott J. Sample Size Estimation in Single-Arm Clinical Trials with Multiple Testing Under Frequentist and Bayesian Approaches. *Journal of Biopharmaceutical Statistics*. 2012; 22(4):819-35. doi: 10.1080/10543406.2012.676585.
118. Wittes J. Sample Size Calculations for Randomized Controlled Trials. *Epidemiologic Reviews*. 2002; 24(1).
119. Collette L. Modelling survival data in medical research. 2nd ed. Boca Raton, Fla: Chapman & Hall/CRC; 2003.

120. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 3rd ed. New York, NY: Springer; 2012.
121. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in medicine*. 1982; 1(2):121–9. doi: 10.1002/sim.4780010204.
122. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. *Biometrics*. 1983; 39(2):499-503.
123. Bland JM, Altman DG. Multiple Significance Tests: The Bonferroni Method. *British Medical Journal*. 1995; 310(6973).
124. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*. 2007; 99(13):1036-43. doi: 10.1093/jnci/djm022.
125. Wang S-J, Hung HMJ, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biometrical journal*. 2009; 51(2):358-74. doi: 10.1002/bimj.200900003.
126. Alosch M, Huque MF, Alosch M, Huque MF. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*. 2009; 28(1):3.
127. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials*. 2010; 31(6):647.
128. Song Y, Chi GYH, Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine*. 2007; 26(19):3535.
129. Chang M. Adaptive design method based on sum of p-values. *STATISTICS IN MEDICINE*. 2007;26(14).
130. Dimairo M, Boote J, Julious SA, Nicholl JP, Todd S. Missing steps in a staircase: a qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials*. 2015; 16(1):430.

131. Spira A, Edmiston KH. Clinical trial design in the age of molecular profiling. *Methods in molecular biology*. 2012; 823:19-34.

132. Medline Plus basic course manual 2012. [accessed on 15 September 2016]. Available online:  
[http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0ahUKEwjS7\\_OmodvJAhWGVhQKHZr0AZMQFggdMAA&url=http%3A%2F%2Fbma.org.uk%2F-%2Fmedia%2Ffiles%2Fpdfs%2Fabout%2520the%2520bma%2Flibrary%2Fmedline%2520plus%2520basic%2520course%2520manual%25202012.pdf&usg=AFQjCNGFxcWiS11CJsroeeIETAWjW0neUA](http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0ahUKEwjS7_OmodvJAhWGVhQKHZr0AZMQFggdMAA&url=http%3A%2F%2Fbma.org.uk%2F-%2Fmedia%2Ffiles%2Fpdfs%2Fabout%2520the%2520bma%2Flibrary%2Fmedline%2520plus%2520basic%2520course%2520manual%25202012.pdf&usg=AFQjCNGFxcWiS11CJsroeeIETAWjW0neUA).

## Chapter 4. Online tool to help develop personalized medicine (BiGTeD)

---

### 4.1. Introduction

---

In this chapter we present an online tool to provide guidance on designing biomarker-guided clinical trials. The literature review work presented in Chapter 2 and 3 of this thesis identified large variability between authors in terms of the terminology used and descriptions of the different biomarker-guided trial designs, which has resulted in significant ambiguity and confusion amongst those trying to interpret and implement the designs. The review also revealed that navigating the literature to gain an understanding of which trial design to implement in a given situation, and the practical implications of doing so, are difficult. Hence our online tool ‘Biomarker-Guided Trial Designs’ (BiGTeD), which is openly and freely available mirrors the findings of the literature review, but in a much more accessible format. To our best knowledge there is no web resource aimed at giving easy access to key information related to each different biomarker-guided clinical trial designs and providing a truly interactive tool allowing for the optimal design in a given setting to be identified.

The decision to develop the tool stemmed from feedback by attendees of conferences and meetings where the literature review work was presented, which suggested that there was a real need for information on the different trial designs to be available in an easily accessible and user-friendly format. The work was also presented at a local North West Hub for Trials Methodology Research meeting at which several members of the Hub’s Stratified Medicine Working Group were present, and it was agreed that having the information identified in the literature review available via a user-friendly web resource would be extremely useful and would increase accessibility to stakeholders. We hope that BiGTeD will improve the understanding of biomarker-guided trial designs and provide much-needed guidance on their implementation.

## 4.2. Key features

BiGTeD is accessible via the following link, [www.BiGTeD.org](http://www.BiGTeD.org), and was developed using Microsoft Expression Web 4. Expression Web is a full-featured professional tool for designing, developing, and publishing compelling, feature-rich websites that conform with web standards.

Since we wished to make the tool as user-friendly as possible, a key feature of the tool is the inclusion of an interactive graphical representation of each trial design, which allows the design and flow of patients through the trial to be easily visualized. Graphical representations have been standardised in such a way that helps to highlight both the similarities and differences across different trial designs. Different colours are used to identify different stages of each design and components of the graphic where additional information is available have been highlighted using a red colour to be easily identifiable. To display this additional information, the user must simply hover the cursor over the relevant component, and key information relating to that component will appear as a 'pop-up' box, e.g. see Figure 4.1.

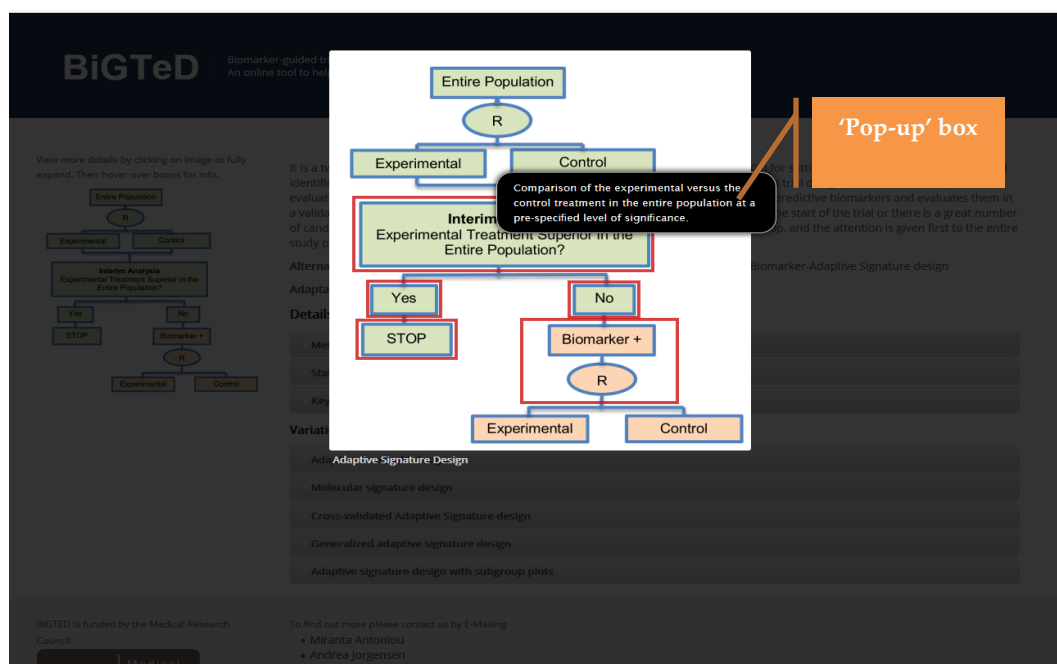


Figure 4.1. 'Pop-up' box illustration.

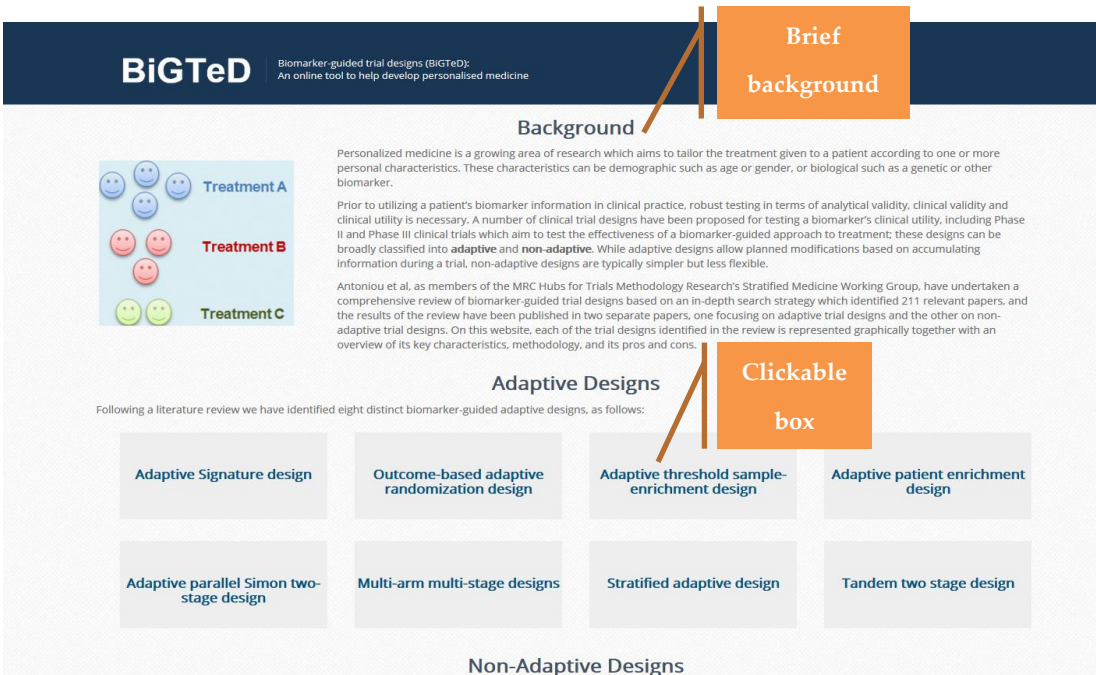


To supplement each graphical representation, key details about the design e.g. the methodology and statistical and practical considerations are provided in more extensive text. Links to references relevant to each trial design identified in our comprehensive and in-depth literature review are also provided so that users can refer to them for further information as required.

### 4.3. User interface

#### 4.3.1. Homepage

The online tool’s homepage includes a brief background to the field of personalized medicine and the utility of biomarker-guided clinical trials. Additionally, a clickable box is provided for each individual trial design, eight for adaptive and five for non-adaptive designs. A snapshot of the homepage is given in Figure 4.2. On clicking on a clickable box, a new webpage is opened which includes the aforementioned graphical representation and key details. For example, if the user clicks on the ‘Adaptive Signature design’ button, the webpage as illustrated in Figure 4.3 opens. It is possible to return to the homepage by simply clicking on the breadcrumb ‘Home’ at the top of the page.



**Figure 4.2.** Online tool’s homepage

### 4.3.2. Design-specific webpages: Adaptive Designs

An example of a design-specific webpage for an adaptive trial is provided in Figure 4.3 which corresponds to the webpage of the Adaptive Signature design. In this figure we can see a shrunk version of the graphic as well as key information about the design, and variations of the design identified in the literature.

The screenshot shows the BiGTED website interface. At the top, the header reads 'BiGTED Biomarker-guided trial designs (BiGTED): An online tool to help develop personalised medicine'. Below this, the page title is 'Adaptive Signature design'. The main content area is divided into several sections: a flowchart on the left, a descriptive paragraph, 'Alternative names', 'Adaptations', 'Details', and 'Variations'. The flowchart illustrates a two-stage process starting with 'Entire Population' (R), branching into 'Experimental' and 'Control' groups. An 'Interim Analysis' box asks 'Experimental Treatment Superior in the Entire Population?'. If 'Yes', it leads to 'STOP'. If 'No', it leads to 'Biomarker +', which then branches into 'Experimental' and 'Control' groups. An orange box labeled 'Shrunk version of the graphic' points to the flowchart. Another orange box labeled 'Key information' points to the 'Details' section, which lists 'Methodology', 'Statistical/Practical considerations', and 'Key references'. The 'Variations' section lists several design types: 'Adaptive threshold design', 'Molecular signature design', 'Cross-validated Adaptive Signature design', 'Generalized adaptive signature design', and 'Adaptive signature design with subgroup plots'. The footer includes funding information from the Medical Research Council and contact details for Miranta Antoniou, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona.

Figure 4.3. Example of the webpage of a distinct adaptive design

On each page a shrunk version of the graphical representation is displayed, which the user can click on to see an expanded version as illustrated in Figure 4.4. The green coloured components of the graphic represent the first stage of the design, whilst, the orange coloured components correspond to the second stage. For each component which is highlighted with a red box, further information about that component of the trial can be displayed if the user hovers over it with the cursor, as illustrated in Figure 4.5. The expanded version can be closed by clicking anywhere on the blackened background.

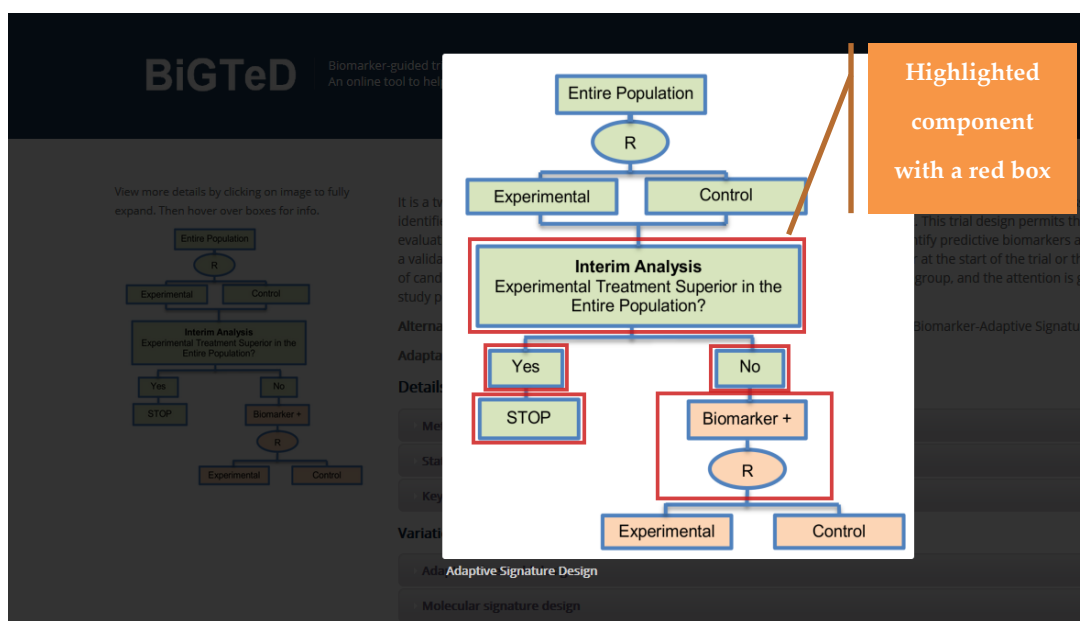


Figure 4.4. Example of an expanded version of an adaptive trial graphic

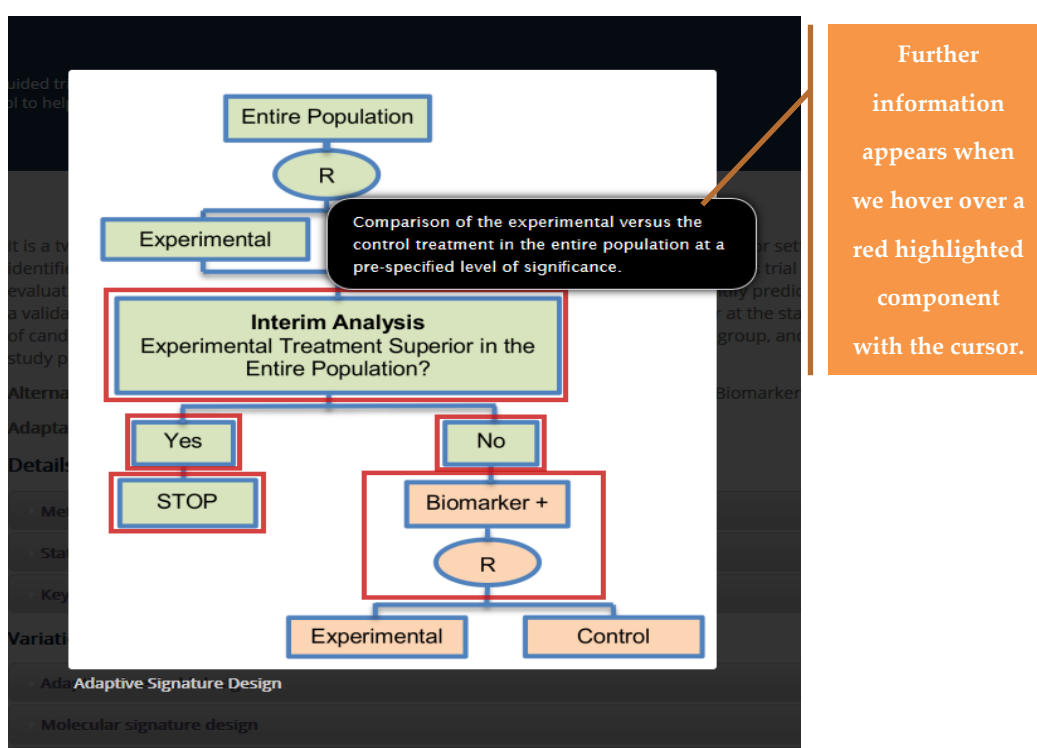


Figure 4.5. Example of an expanded version of an adaptive design graphic with the 'pop-up' box showing further information

Each design-specific webpage also includes a section entitled 'Details', which includes three clickable boxes. Clicking on the first two clickable boxes will reveal key information relating to the methodology (Figure 4.6) or statistical and practical information respectively (Figure 4.7), whilst clicking on the third clickable box will

reveal key references about the trial as identified in our comprehensive literature review given in Chapter 2 (Figure 4.8).

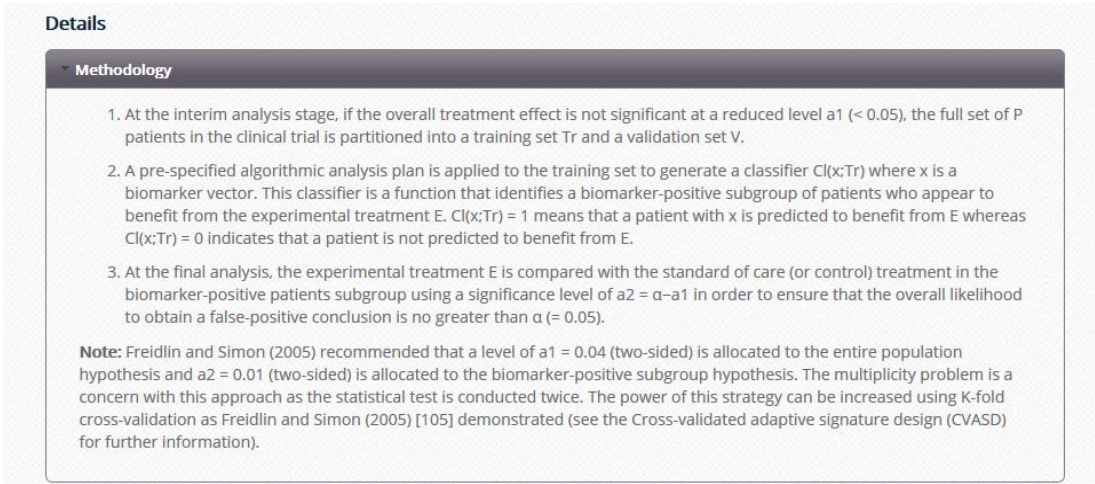


Figure 4.6. Methodology information in the ‘Details’ section of an adaptive design graphic.

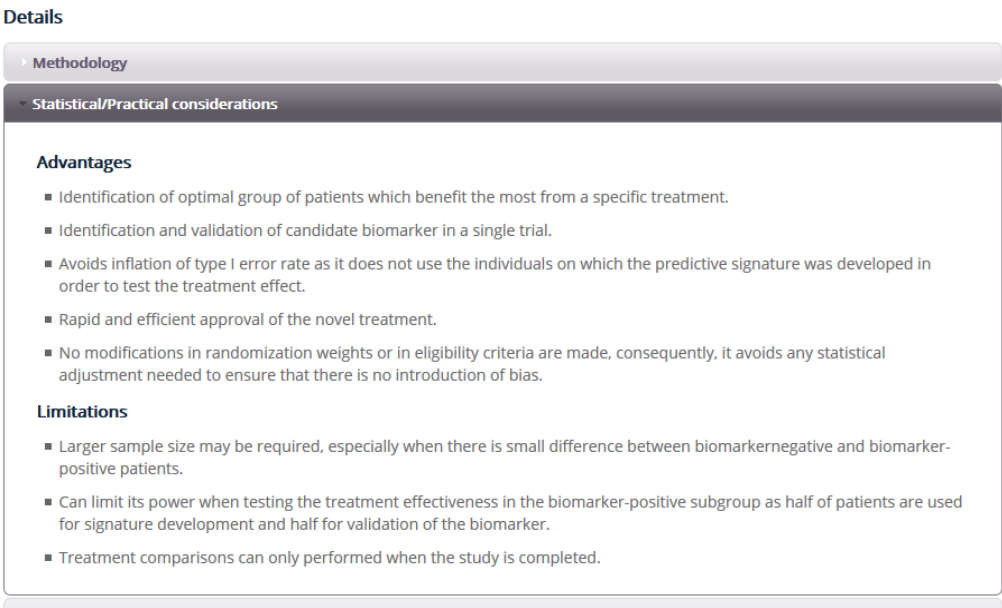
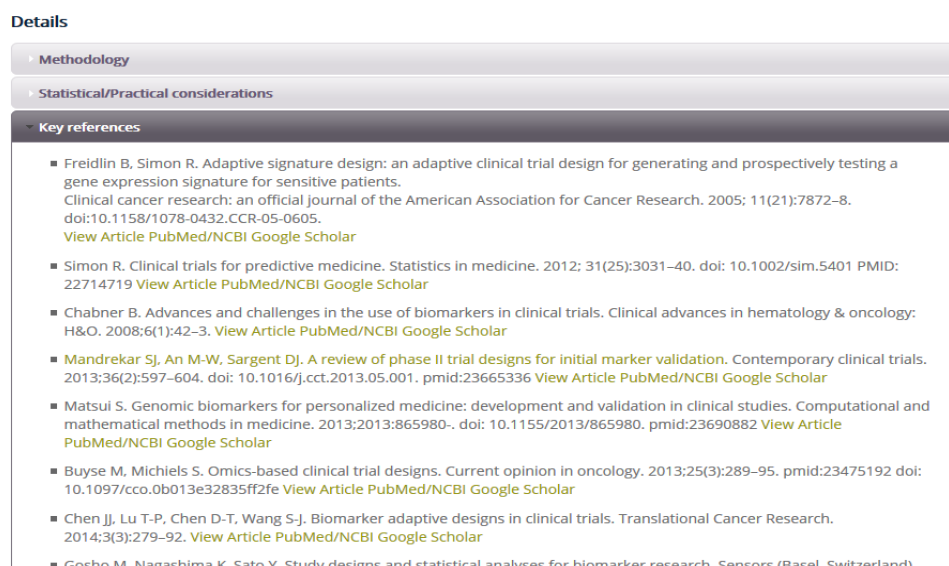
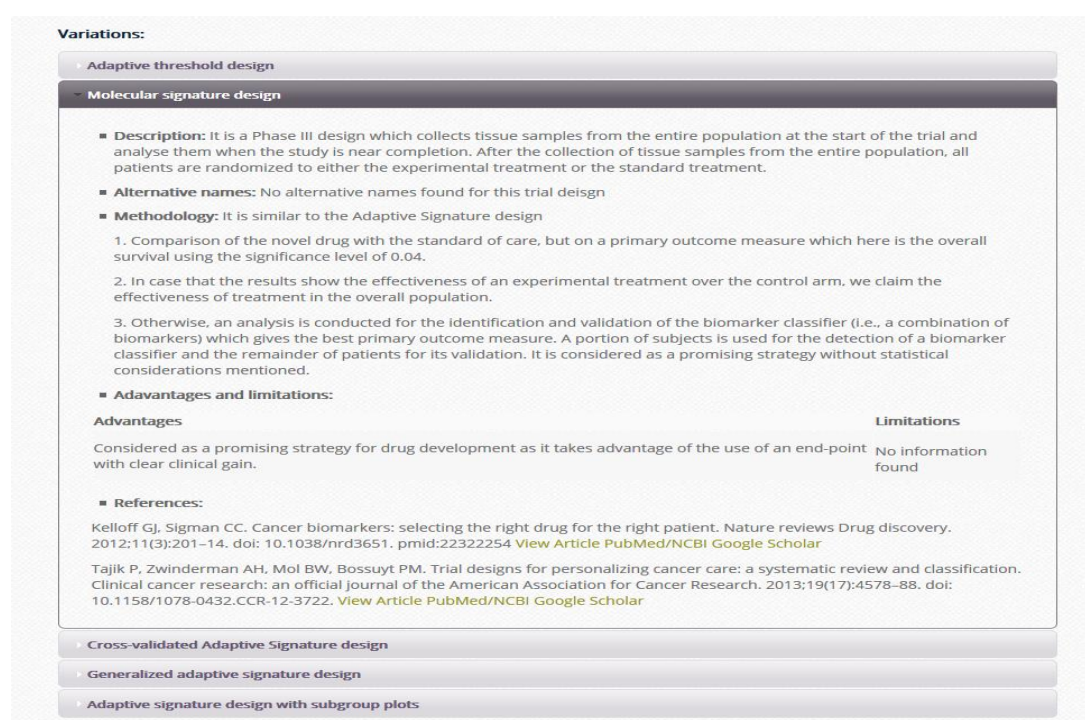


Figure 4.7. Statistical and practical information in the ‘Details’ section of an adaptive design graphic.



**Figure 4.8.** Key references in the ‘Details’ section of an adaptive design graphic.

Each design-specific webpage also includes a section entitled ‘Variations’ which includes several clickable boxes, each representing variations of the trial design under consideration, again as identified in our comprehensive literature review given in Chapter 2. Clicking on a clickable box reveals detailed information and key references for the variation (please see Figure 4.9 as an example).

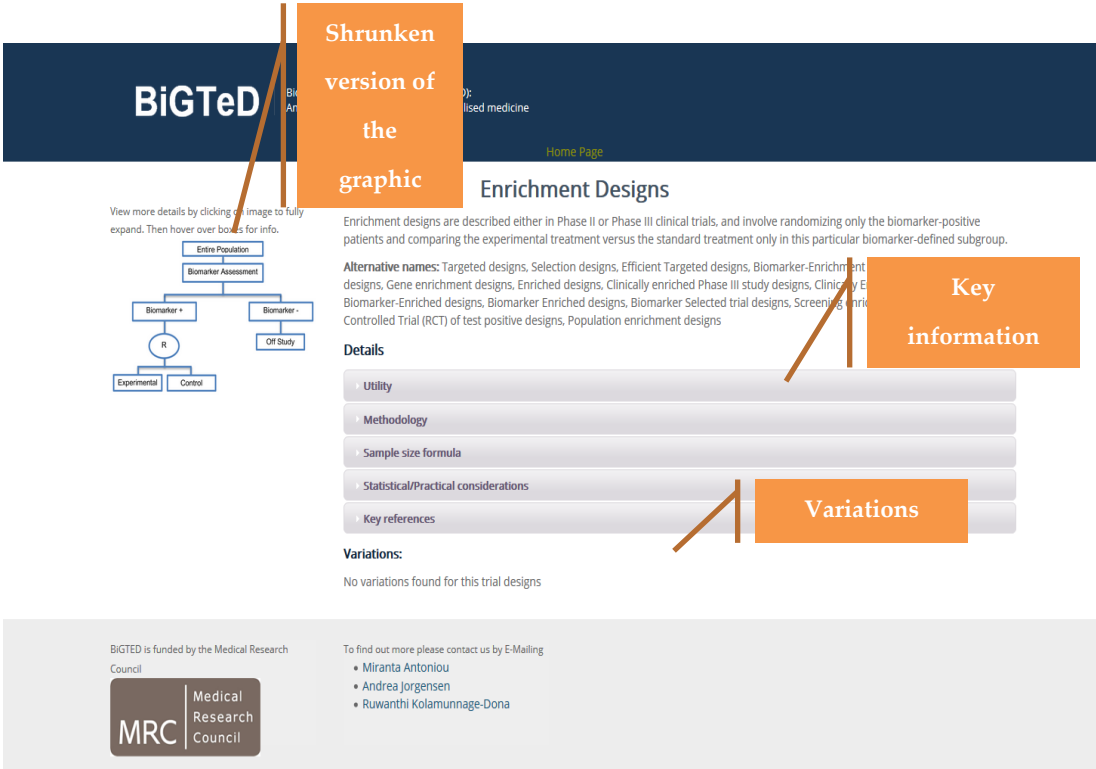


**Figure 4.9.** ‘Variations’ section of an adaptive design graphic

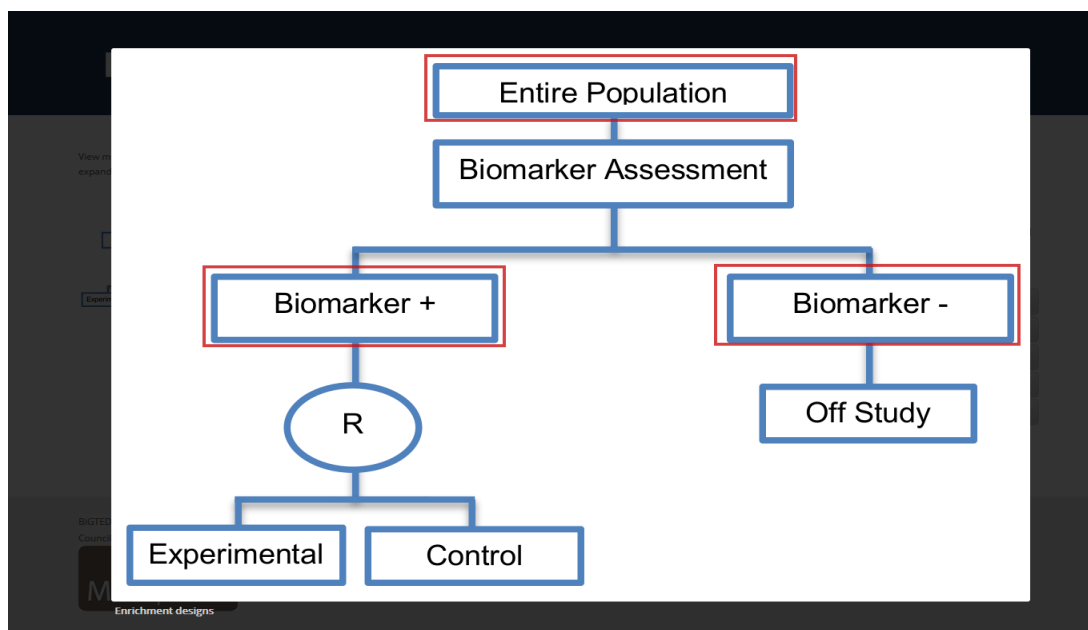


### 4.3.3. Design-specific webpages: Non-Adaptive Designs

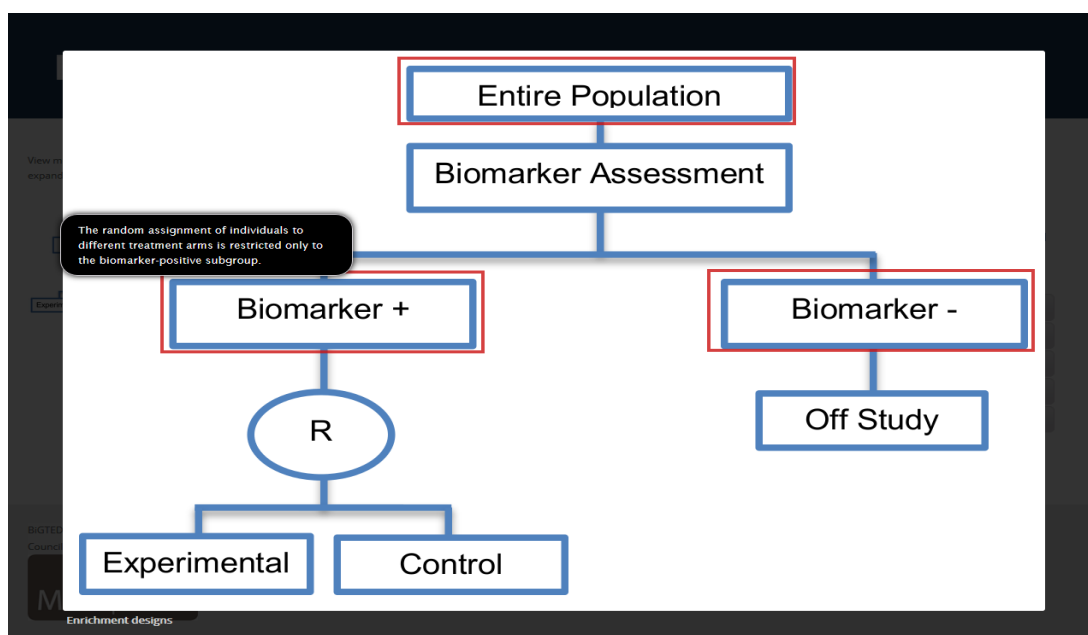
In a similar way to the adaptive designs, the main page of each individual non-adaptive design is composed of a shrunk version of a graphical representation of the design, key information and variations to the design. A snapshot of the main page of one of the five distinct non-adaptive designs (the Enrichment design) is given in Figure 4.10. Similarly to the adaptive designs, the user can see an expanded version of the shrunk graphical representation by clicking on the graphic (see Figure 4.11). For each component which is highlighted with a red box, further information about that component of the trial can be displayed if the user hovers over it with the cursor, as illustrated in Figure 4.12. The expanded version can be closed by clicking anywhere on the blackened background.



**Figure 4.10.** Example of the webpage of a distinct non-adaptive design



**Figure 4.11.** Example of an expanded version of a shrunk non-adaptive design graphic



**Figure 4.12.** Example of an expanded non-adaptive design graphic with a 'pop-up' box showing further information

For the non-adaptive designs, the 'Details' section includes five clickable boxes. The first four can be clicked on to find out further details about the utility (Figure 4.13), methodology (Figure 4.14), sample size calculation (Figure 4.15), and statistical/practical considerations (Figure 4.16) for the design respectively, whilst clicking on the fifth will reveal key references (Figure 4.17) for the design as identified in our literature review presented in Chapter 3. As for the adaptive designs, each non-

adaptive design-specific webpage also includes a section entitled ‘Variations’ which includes several clickable boxes, each representing a variation of the trial design under consideration, again as identified in our comprehensive literature review given in Chapter 3.

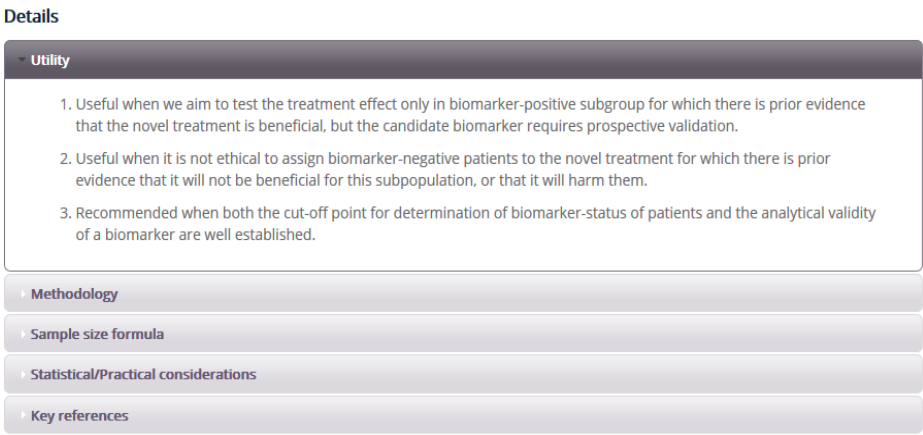


Figure 4.13. Utility information in the ‘Details’ section of a non-adaptive graphic

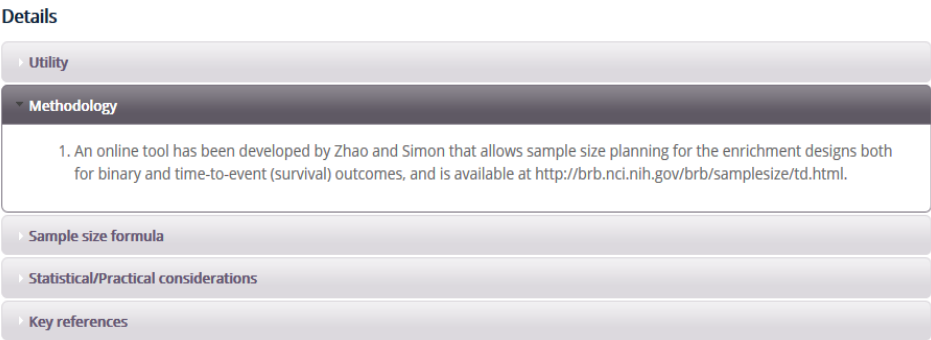


Figure 4.14. Methodology information in the ‘Details’ section of a non-adaptive graphic



## Details

### Utility

### Methodology

### Sample size formula

$$E(D_{\text{enrichment}}) = \frac{nT\lambda_i}{2(\lambda_i + \varphi)} \left\{ 1 - \frac{e^{-(\lambda_i + \varphi)\tau}}{(\lambda_i + \varphi)\tau} [1 - e^{-(\lambda_i + \varphi)\tau}] \right\}$$

$E(D_{\text{enrichment}})$  is referred to the expected number of events per treatment arm (time-to-event outcome),  $i$  corresponds to either the experimental or the control treatment group, **1:1** ratio between the two treatment arms (experimental:control) is assumed,  $\lambda$  corresponds to the event hazard rate,  $\varphi$  is the loss to follow-up rate,  $T$  denotes the accrual time, patients enter the trial according to a Poisson process with rate  $n$  per year over the accrual period of  $T$  years,  $\tau$  corresponds to the follow-up period.

$$D_{\text{enrichment}} = 4 \left[ \frac{(z_{\alpha/2} + z_{\beta})^2}{\log \theta_1} \right]^2$$

$D_{\text{enrichment}}$  is referred to the required total number of events (time-to-event outcome), **1:1** ratio between the two treatment arms (experimental:control) is assumed,  $z_{\alpha/2}$ ,  $z_{\beta}$  denote the upper  $\alpha/2$ - and upper  $\beta$ -points respectively of a standard normal distribution,  $\alpha$  and  $\beta$  denote the assumed type I error and type II error respectively,  $\theta_1$  denotes the assumed hazard ratio between the two treatment groups (control vs experimental) in the biomarker-positive subgroup.

$$N_{\text{enrichment/arm}} = 2\bar{p}_Q (1 - \bar{p}_Q) \left[ \frac{(z_{\alpha/2} + z_{\beta})^2}{(p_A^Q - p_B)} \right]^2$$

$N_{\text{enrichment/arm}}$  is referred to the required number of patients per treatment arm (binary outcome), **1:1** ratio between the two treatment arms (experimental:control) is assumed,  $p_A^Q$  and  $p_B$  are the response probabilities in the experimental and control

Figure 4.15. Sample size formulae in the 'Details' section of a non-adaptive graphic

## Details

### Utility

### Methodology

### Sample size formula

### Statistical/Practical considerations

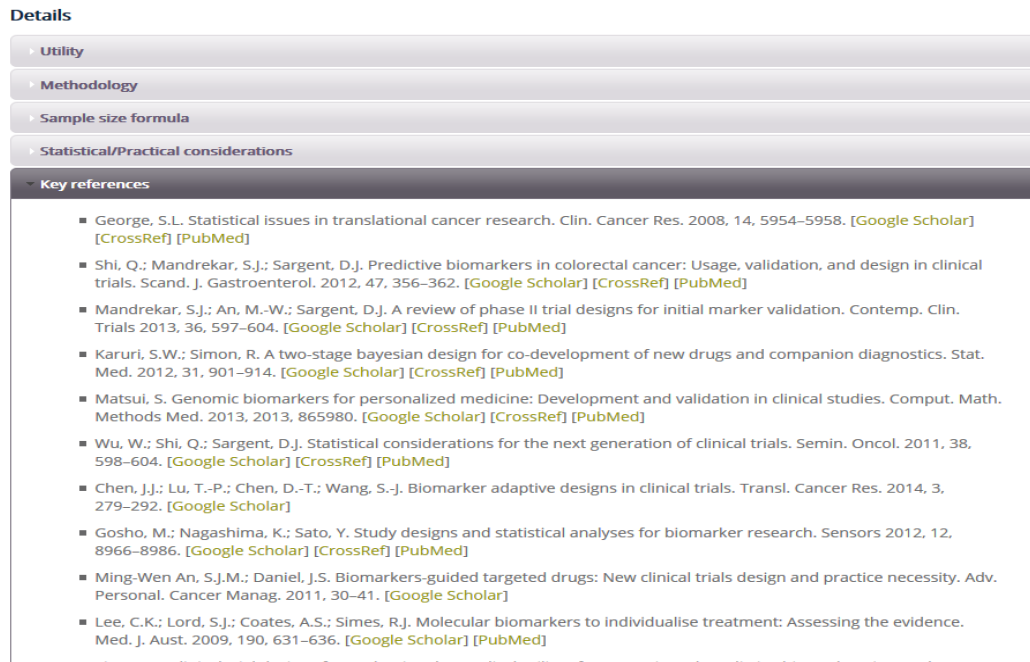
#### Advantages

- Evaluate the effect of the experimental treatment in the biomarker-positive subgroup in a simple and efficient way.
- Provide clear information about whether the novel treatment is effective for the biomarker-positive subgroup, thus these designs can identify the best treatment for these patients and confirm the usefulness of the biomarker.
- Reduced sample size as the assessment of treatment effect is restricted only to biomarker-positive subgroup. Therefore, if the selected biomarker is "biologically correct" and reliably measured, the used enrichment strategy could result in a large saving of randomized patients.
- Enable rapid accumulation of efficacy data.
- Avoid potential dilution of the results due to the absence of biomarker-negative patients. For example, if the design had included the biomarker-negative population and the biomarker positive prevalence was low as compared to the biomarker negative prevalence, then the estimation of the overall treatment effectiveness could be diluted as it would be driven by the biomarker-negative subgroup.
- Can be attractive in terms of speed and cost, meaning that patients are provided with tailored treatment sooner.

#### Limitations

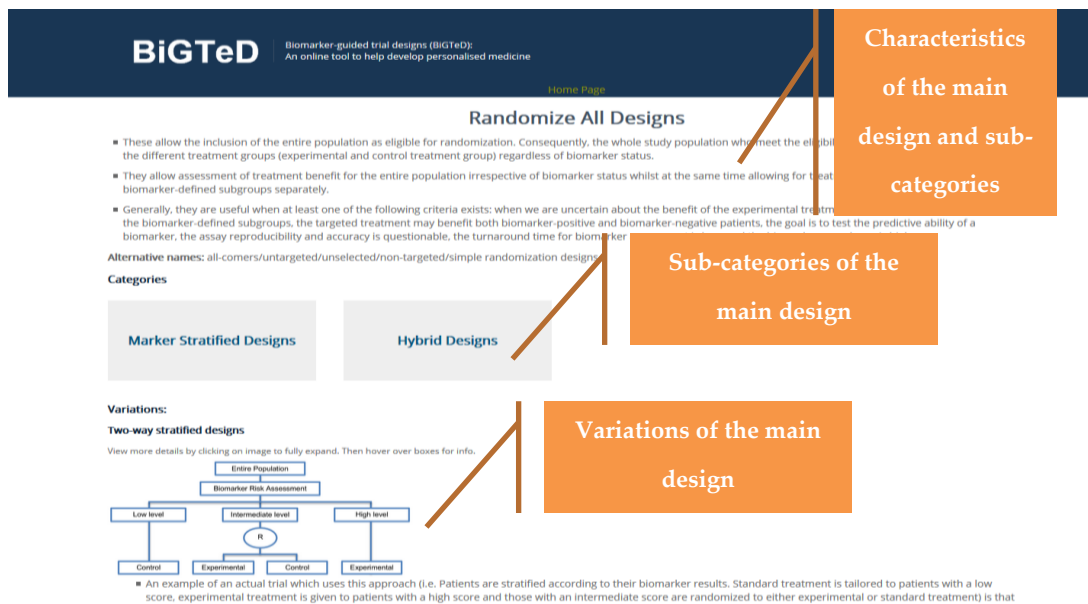
- Do not assess whether the experimental treatment benefits the biomarker-negative patients, thus we cannot obtain information about this subgroup.
- Unable to demonstrate whether the targeted treatment is beneficial in the entire study population.
- Do not inform us directly about whether the biomarker is itself predictive because the relative treatment efficacy may be the same in the unevaluated biomarker-negative patients. Since these designs only enrol a subgroup of patients, they do not allow for full validation of the marker's predictive ability. For full validation, a trial would need to randomize all patients in order to test for a treatment-biomarker interaction.
- Researchers should carefully consider whether or not to follow this strategy as it may be of limited value due to the exclusion of biomarker-negative patients. It may be that the entire population could benefit from the experimental treatment equally irrespective of biomarker status, in which case enrolling only the biomarker-positive patients will result in slow trial accrual, increase of expenses and unnecessary limitation of the size of the indicated patient population.
- Concern over an ethical problem as we cannot include individuals in a clinical trial if it is believed that the treatment is not effective for them, as raised by the US Food and Drug Administration (FDA). It was based on the fact that the experimental treatment can only be approved for a particular biomarker-defined subpopulation (i.e., biomarker-positive patients) if a companion diagnostic test is also approved and how the test can be approved if the Phase III trial does not show that the

Figure 4.16. Statistical and practical information in the 'Details' section of a non-adaptive graphic



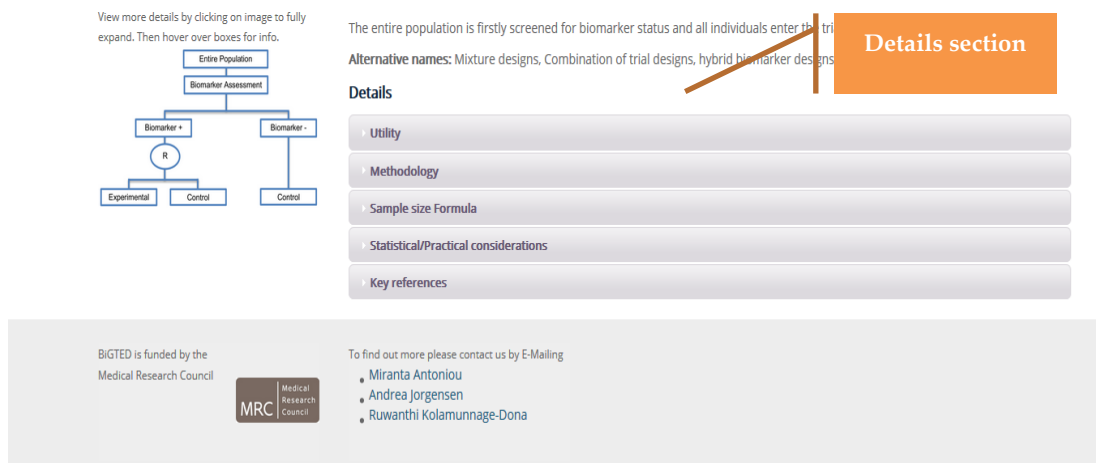
**Figure 4.17.** Key references in the ‘Details’ section of a non-adaptive graphic

Some non-adaptive designs are divided into sub-categories which share the same characteristics, e.g. for the Randomize-all Designs’, there are two sub-categories, the so-called ‘Marker Stratified Designs’ and the ‘Hybrid Designs’. In these cases, the design-specific webpage has a different layout as illustrated in Figure 4.18. More precisely, first we present the characteristics which are relevant to both the main design and the sub-categories (i.e. Randomize-all Designs) and all the alternative names found in the literature (see Chapter 3). Next, the different sub-categories included in the main category are given (see grey clickable boxes in Figure 4.18). When we click on those boxes then key information and graphical representations relating to the sub-category specifically are presented. Apart from the sub-categories, variations of the main design are also provided.



**Figure 4.18.** Design-specific webpage of a non-adaptive design divided in different sub-categories.

The sub-categories of each main design follow similar structure used in the distinct designs. Hence, a 'Details' section is provided. A snapshot of one sub-category of the 'Randomize All Designs', the so-called 'Hybrid Designs' can be seen in Figure 4.19.



**Figure 4.19.** Design-specific webpage of a sub-category of a non-adaptive design

## 4.4. Discussion

To conclude, in the current chapter we presented BiGTed which provides a unique resource where details of all biomarker-guided trial designs can be easily explored, and will be a valuable tool for those involved in planning and

implementing trials in personalized medicine. A key feature is the inclusion of the interactive graphical representations of the trial designs. These clear graphics allow the design and flow of patients through the trials to be visualized, whilst having them standardized in terms of structure and colour coding allows the different designs to be easily compared. On presenting BiGTeD at conferences and meetings, we have received very positive feedback from academics, industry and research councils on its potential utility in guiding and encouraging the adoption of the most effective and appropriate design. It is anticipated that users could use and cite the graphical representations of BiGTeD in trial protocols and funding applications to aid in describing the trial design. We aim to continue to raise awareness of the website via presentations at conferences and meetings.

Although initially BiGTeD reflects the information identified in the reviews reported in chapters 2 and 3, future plans include extending the tool to introduce an interactive element whereby certain aspects of a trial's setting can be input with suggested optimal trial designs for that setting being output, as well as the introduction of sample size/power calculators to support each design. To inform such plans a workshop will be arranged, attended by experts in the field, with the aim of gaining consensus in terms of developing BiGTeD into a truly interactive website for the purpose of choosing and designing a biomarker-guided trial, what might be deliverable on such an interactive website, and what further methodological work may be required prior to making the tool truly interactive in terms of suggesting an appropriate trial design. Potential participants for the workshop will include members of the Hub Network's Stratified Medicine Working Group and the Adaptive Trials Working Group. Additional participants, including clinical trial unit representatives, will also be invited to ensure that the needs of the end-user are addressed when deciding on how to extend the tool.

In the next chapter (Chapter 5) we will explore the characteristics of one of the popular non-adaptive trial design and an adaptive approach of it with the aim to assess the efficiency of the study related to the cost and time of the trial in general.

We will showcase the way that simulation techniques can be performed for the calculation of the total study power.

## Chapter 5. Fixed and adaptive Parallel Subgroup-Specific design for survival outcomes: power and sample size

---

### 5.1. Introduction

---

In Chapter 1, we introduced the growing field of personalized medicine [1]. Following on from Chapters 2 and 3, we presented a number of biomarker-guided clinical trial designs [2-3] and showed BiGTeD in Chapter 4 which is informed by the literature review discussed in previous chapters.

Whilst such designs have been given significant attention in the literature, there are many challenges associated with their design, analysis and practical application, which need to be explored further and better understood. Key challenges include powering the study adequately, controlling the false-positive rate, and applying appropriate stopping probabilities.

In the current chapter we focus on the “Parallel Subgroup-Specific” design [4-6] which can be explored to discuss the above challenges. The parallel subgroup-specific design is used to test the clinical hypothesis of treatment effect, evaluating the effect of the experimental treatment relative to a control treatment in both a biomarker-negative and a biomarker-positive subgroup separately. We also consider an adaptive version of the design, splitting the trial into two-stages with the aim of stopping the study early for either a positive or negative outcome. In this adaptive version, the first stage involves an interim analysis after the pre-specified percentage of events are achieved and a decision is made whether to stop the trial early for efficacy or futility, or to continue to the second stage of the study. Futility and efficacy are assessed by comparing the p-value of the observed test statistics produced at each stage of the design with pre-specified stopping boundaries. The role of an interim analysis in a clinical trial design is important as it might allow the experimental

treatment to be made available earlier in case of positive results. We have conducted several simulation studies to evaluate a variety of scenarios.

Our aim was to explore one of the most popular designs in personalized medicine as it came out from our in depth literature reviews in Chapter 2, 3, to the adaptive version of it in terms of their statistical characteristics (e.g. study power, expected sample size) and showcase the way that simulations can be performed. Hence, in the current chapter, we explore a fixed versus an adaptive approach in a popular biomarker-guided clinical trial setting. This to our best knowledge has not been investigated yet.

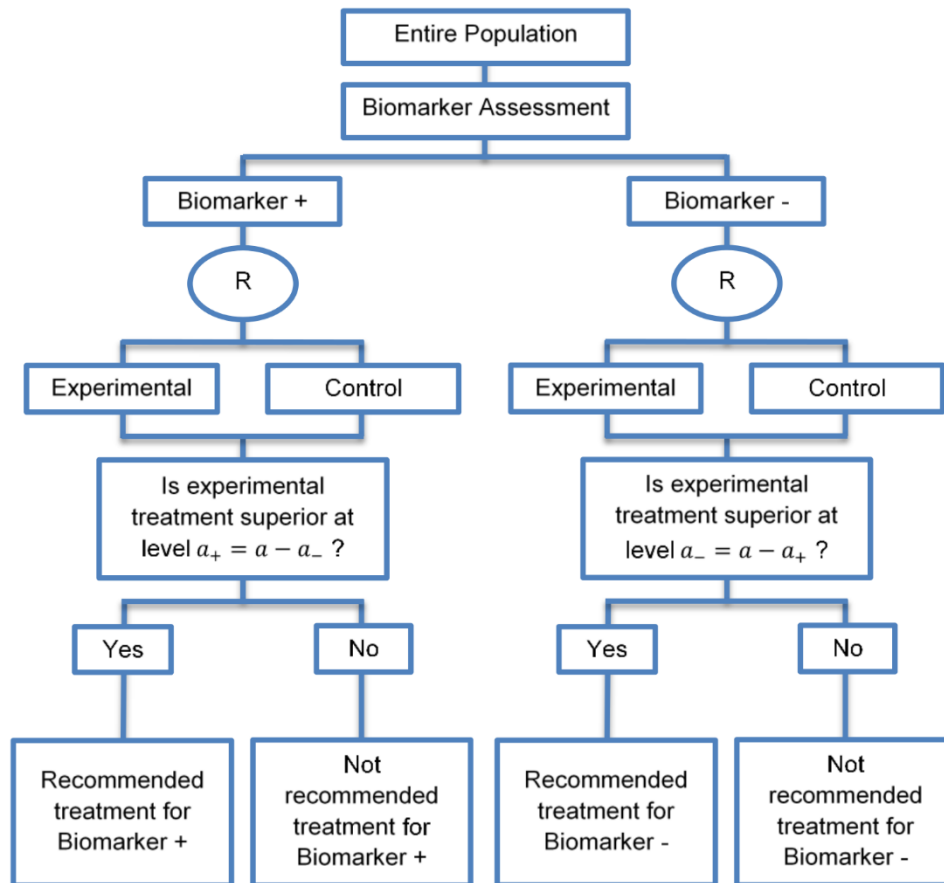
## 5.2. Methods and Findings

---

### 5.2.1. Parallel Subgroup-Specific design

---

The parallel subgroup-specific design, a modified version of the marker stratified design, allows for the evaluation of treatment effects separately in the biomarker-positive subgroup and biomarker-negative subgroup at the same time [3]. Whilst the marker-stratified design uses the overall significance level  $\alpha$  for each biomarker-defined subgroup separately, the parallel subgroup-specific design controls the overall type I error rate by splitting the overall significance level  $\alpha$  between the two biomarker subgroup tests such that  $\alpha = \alpha_- + \alpha_+$  [3]. A graphical illustration of this strategy is given in Figure 5.1.



**Figure 5.1.** Parallel Subgroup-Specific design. “R” refers to randomization of patients.  $\alpha$  refers to the overall significance level between the two biomarker subgroup tests such that  $\alpha = \alpha_- + \alpha_+$ .

All patients are screened for biomarker status (biomarker positive or biomarker negative) and then randomized to the experimental or control treatments in the two biomarker subgroups. Therefore, biomarker status acts as a stratification factor. Consequently, the trial is made up of four arms, i.e. biomarker-positive patients receiving either the experimental or the control treatment and biomarker-negative patients receiving either the experimental or the control treatment. A test for treatment effectiveness (Experimental treatment vs Control treatment) can therefore be performed in each biomarker-defined subgroup separately.

Where a trial’s primary outcome is time to some specified event (e.g. time to death), the hypotheses being tested in the two biomarker subgroups if one assumes exponentially distributed times can be defined as follows:



i) **Hypothesis being tested (case of two-sided test) in the biomarker negative**

**subgroup:**  $H_{0,biom-}: \log(\theta_-) = 0$ , where

$$\theta_- = HR_{biom-} = \frac{\lambda_{E-}}{\lambda_{C-}}$$

denotes the hazard ratio, and  $\lambda_{E-}$  and  $\lambda_{C-}$  are the rate parameters of an exponential distribution for biomarker-negative patients receiving experimental treatment and control treatment respectively and

ii) **Hypothesis being tested (case of two-sided test) in the biomarker positive**

**subgroup:**  $H_{0,biom+}: \log(\theta_+) = 0$ , where

$$\theta_+ = HR_{biom+} = \frac{\lambda_{E+}}{\lambda_{C+}}$$

refers to the hazard ratio, and  $\lambda_{E+}$  and  $\lambda_{C+}$  are the rate parameters of an exponential distribution for biomarker-positive patients receiving experimental treatment and control treatment respectively.

#### 5.2.1.1. Sample size calculation for time-to-event-outcomes

---

For the purpose of undertaking power calculations for this design, we assume that the treatment effect will be tested using the log-rank test. The total number of events required for the parallel subgroup-specific design can be calculated by adding up the number of events required in each biomarker-defined subgroup. Following Mandrekar and Sargent (2009) [7], we assume 1:1 randomization, and therefore the required number of events for each biomarker-defined subgroup can be calculated by

$$D_j = 4 \frac{(z_{a_j} + z_{\beta})^2}{[\log(\theta_j)]^2} \quad (5.1)$$

where  $j$  denotes either the biomarker positive subgroup ( $j = +$ ) or the biomarker negative subgroup ( $j = -$ ),  $z_{a_j}$ ,  $z_{\beta}$  denote the upper  $a_j$ - and upper  $\beta$ -points

respectively of a standard normal distribution and the required total number of events can be calculated by

$$D = D_- + D_+ = 4 \frac{(z_{a_-} + z_\beta)^2}{[\log(\theta_-)]^2} + 4 \frac{(z_{a_+} + z_\beta)^2}{[\log(\theta_+)]^2} \quad (5.2)$$

where  $a_-$  and  $a_+$ , denote the type I error rates for biomarker-negative and biomarker-positive subgroup respectively such that  $a_- + a_+ = a$ , and  $a$  is the nominal significance level (if one-sided e.g.,  $a = 0.025$  in our case) and  $\beta$  corresponds to the type II error rate (it is common across the two subgroups). One-sided significance levels are used in situations where an alternative hypothesis specifies that the treatment benefit in the experimental group is greater than that of the control group. In case that a two-sided  $a$  is used (e.g.,  $a = 0.05$ ), then the required total number of events can be calculated by

$$D = D_- + D_+ = 4 \frac{(z_{a_-/2} + z_\beta)^2}{[\log(\theta_-)]^2} + 4 \frac{(z_{a_+/2} + z_\beta)^2}{[\log(\theta_+)]^2}.$$

When more than one hypothesis for the assessment of experimental treatment efficacy is being tested, it is important to control the familywise error rate (FWER) by adjusting for multiplicity of testing to ensure that the probability to commit at least one Type I error does not exceed the nominal significance level. To achieve this, a conservative Bonferroni correction method is often used where  $a$  is allocated between the test for the biomarker-negative subgroup and the test for the biomarker-positive subgroup either equally (i.e.  $a/2$ ) or unequally, meaning that the significance levels assigned to each biomarker defined-subgroup then add up to the total significance level  $a$ .

The calculation of the total sample size needed for this study is based on the total number of events and the probability that a subject will get an event prior to the end of the study [8]. Therefore, the sample size required for subgroup  $j$  is,

$$N_j = \frac{D_j}{Pr_j(event)} \quad (5.3)$$

where  $j$  refers to the biomarker-defined subgroup,  $Pr_j(event)$  corresponds to the probability of observing an event in biomarker subgroup  $j$  which can be calculated by

$$Pr_j(event) = \pi_E Pr_{Ej}(event) + \pi_C Pr_{Cj}(event),$$

with

$$\pi_E = \frac{R}{R+1} \text{ and } \pi_C = \frac{1}{R+1}.$$

$\pi_E$  and  $\pi_C$  are the proportions of patients who are randomized to the experimental and control treatment arm respectively.  $R$  is the allocation ratio which is given by the sample size in experimental arm divided by the sample size in control arm. Here we assume equal allocation between treatment arms for each biomarker-defined subgroup [9]; hence  $R = 1$  and  $\pi_E = \pi_C = 0.5$ .  $Pr_{Ej}(event)$  and  $Pr_{Cj}(event)$  are the probabilities of event in the experimental and control treatment arm respectively in subgroup  $j$ . If  $i$  now denotes treatment group (either experimental ( $E$ ) or control ( $C$ )) and if one assumes exponentially distributed times as described in detail in [9], the probability of an event in treatment arm  $i$  of subgroup  $j$  can be calculated by

$$Pr_{ij}(event) = 1 - \frac{1}{\left(\frac{\log(2)}{m_{ij}}\right) \times T_j} \times \left[ e^{-\left(\frac{\log(2)}{m_{ij}}\right) \times \tau_j} - e^{-\left(\frac{\log(2)}{m_{ij}}\right) \times (T_j + \tau_j)} \right] \quad (5.4)$$

where  $T_j$  corresponds to the length in months of the accrual period, during which a homogeneous Poisson entry process is assumed, of the biomarker-defined subgroup  $j$  and  $\tau_j$  corresponds to the follow-up period of the biomarker-defined subgroup  $j$ .  $m_{ij}$  denotes the median survival time of treatment arm  $i$  in biomarker-defined subgroup  $j$  where

$$m_{E-} = \frac{m_{C-}}{\theta_-} \quad (5.5)$$

and

$$m_{E+} = \frac{m_{C+}}{\theta_+}. \quad (5.6)$$

Formula (5.4) could be generalized to arbitrary, continuous survival functions.

Using the sample size of each biomarker-defined subgroup, the corresponding accrual rate (number of patients recruited per month) for subgroup  $j$  is  $ar_j$  which can be calculated by

$$ar_j = \frac{N_j}{T_j}. \quad (5.7)$$

#### 5.2.1.2. Simulation Study 1

---

The scope of this simulation study is to confirm that we can achieve the desirable power in each biomarker-defined subgroups under different simulation settings for a time-to-event outcome. We calculate the required number of events and patients for each biomarker-defined subgroup  $(D_j, N_j)$  from (5.1) and (5.3). Different scenarios are considered by varying hazard ratios  $(\theta_-, \theta_+)$  and significance levels  $(a_-, a_+)$ . In our simulation study, we assume that the biomarker-negative patients have a worse treatment outcome as compared to the biomarker-positive subgroup. We assume outcome to be an adverse effect such as time to death and so the assumed hazard ratio values  $< 1$  reflects the fact that the experimental treatment is superior to the control treatment in both biomarker subgroups. Further, the lower hazard ratio value assumed for a specific biomarker-defined subgroup reflects a greater treatment effect in that subgroup. Hence, in all scenarios of hazard ratios, we consider higher  $\theta_-$  than  $\theta_+$ . More specifically, four scenarios of hazard ratios, i.e. (i)  $\theta_- = 0.6$  and  $\theta_+ = 0.4$ , (ii)  $\theta_- = 0.7$  and  $\theta_+ = 0.5$ , (iii)  $\theta_- = 0.8$  and  $\theta_+ = 0.6$  and (iv)  $\theta_- = 0.9$  and  $\theta_+ = 0.7$  and three scenarios of significance levels (i)  $a_- = a_+ = 0.0125$ , (ii)  $a_- = 0.015$  and  $a_+ = 0.010$  and (iii)  $a_- = 0.010$  and  $a_+ = 0.015$  are considered. We set the median survival time of biomarker-negative subgroup in Control group ( $m_{C-}$ ) in (5.5) at 5 months and we calculate the corresponding median survival time for the

Experimental group in that subgroup. We set the median survival time of biomarker-positive subgroup in Control group ( $m_{c+}$ ) in (6.6) at 10 months and we calculate the corresponding median survival time for the Experimental group in that subgroup. Additionally, we set the type II error rate at 20%, i.e.  $\beta = 0.2$  in (5.1) which corresponds to 80% power (i.e.  $1 - \beta = 0.8$ ), length of accrual period ( $T_j$ ) in (5.4) at 18 months and length of follow-up period ( $\tau_j$ ) in (5.4) at 12 months for each biomarker-defined subgroup. Study entry times and event times for each biomarker-defined subgroup are generated as described below.

The time of study entry for participants in each biomarker-defined subgroup is modelled with a Uniform distribution for entry times. More precisely, the entry times of patients recruited into the biomarker negative subgroup in the first month are assumed by randomly generating  $ar_-$  (the accrual rate) numbers from  $U \sim Unif[0,1]$  (i.e. randomly splitting the accrual period by the number of recruited patients). Similarly,  $ar_+$  numbers are generated from  $U \sim Unif[0,1]$  to obtain study entry times of patients recruited into the biomarker positive subgroup during the first month. To obtain study entry times for those in the biomarker negative and biomarker positive subgroups during the second month, a further  $ar_-$  and  $ar_+$  numbers respectively are randomly generated from  $U \sim Unif[1,2]$ . The accrual continues until the assumed accrual period  $T_j$ . Thus, in the  $T_j^{th}$  month, study entry times are generated from  $U \sim Unif[T_j - 1, T_j]$ . At the end of the accrual period  $N_+$  and  $N_-$  participants in total have been recruited.

Event times are generated from an exponential distribution assuming hazard rate  $\lambda_{ij}$  for  $j$ th biomarker-defined subgroup receiving treatment  $i$ . The values of  $\lambda_{ij}$  can be determined by

$$S_{ij}(m_{ij}) = \exp(-\lambda_{ij} \times m_{ij}) = 0.5,$$

where  $m_{ij}$  are corresponding median survival times, and  $S_{ij}(m_{ij})$  is the exponential median survival probability for subgroup  $j$  and treatment  $i$ . By solving  $S_{ij}(m_{ij})$  for  $\lambda_{ij}$  gives

$$\lambda_{ij} = \frac{\ln 2}{m_{ij}}.$$

We assume patients are not lost to follow-up during the study, and hence any censoring in both biomarker-defined subgroups is due to the event occurring after a ‘cutoff’ time. The cutoff time refers to the time after study start at which a pre specified number of events  $D_j$  for each biomarker-defined subgroup has been reached. A time  $t_{ij}$  (i.e. sum of accrual time and follow-up time) is generated for each patient, and if  $t_{ij}$  is less than the cutoff time then it is assumed that the event was observed at  $t_{ij}$ , otherwise the patient’s event time is censored at  $t_{ij}$ .

One-sided p-value for treatment effect in each biomarker-defined subgroup are computed using the log-rank test. One-sided p-value are considered because we assume that the treatment benefit in the experimental group is greater than that in the control group.

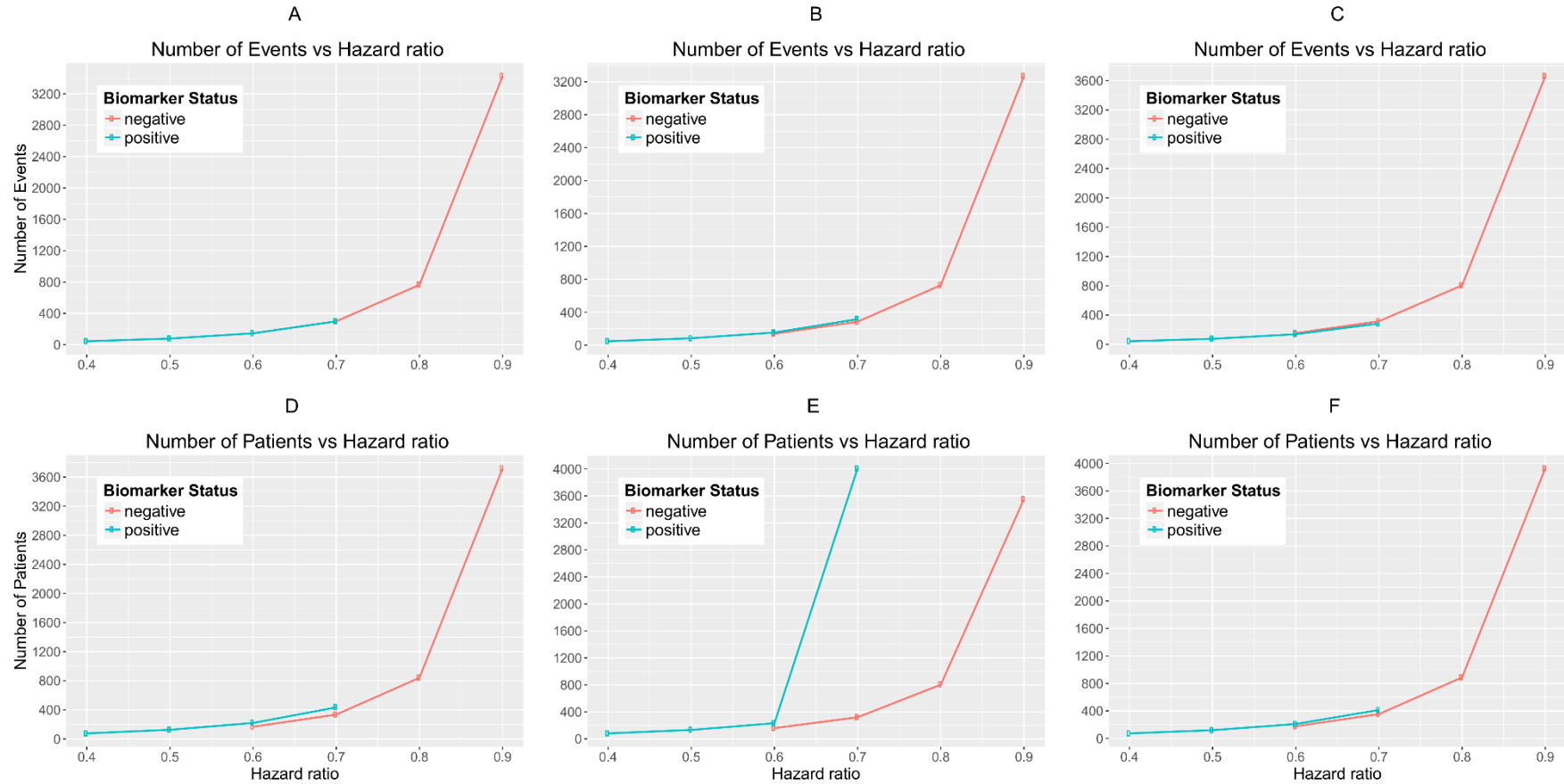
#### *5.2.1.3. Results from simulation study 1*

---

The results are drawn from 10000 iterations. The simulated power of each biomarker-defined subgroup is preserved approximately at 80% across all scenarios of hazard ratios and significance levels. The accrual rates and the number of events and patients to reach the nominal level of power (80%) corresponding to different scenarios of hazard ratios and significance levels are presented in Table C.1 provided in Appendix C. The power for each biomarker-defined subgroup yielded from the simulation study is also presented. Figure 5.2 (A-C) illustrates the required number of events for each biomarker-defined subgroup versus the corresponding hazard ratio for each of the three scenarios of significance levels. Figure 5.2 (D-F) illustrates the required number of patients of each biomarker-defined subgroup versus the corresponding hazard ratio for each of the three scenarios of significance levels. As expected, the number of events and therefore the sample size required for each biomarker-defined subgroup increases with the increase of the corresponding hazard ratio at the same significance level. Furthermore, at each scenario of hazard ratio, we can achieve a smaller sample size and necessary number of events for each

biomarker-defined subgroup with a larger significance level (for example, when  $HR$  scenario (i)  $\theta_- = 0.6$  and  $\theta_+ = 0.4$ , and when  $\alpha_+ = 0.015$ , we achieve the smallest necessary number of events and sample sizes).

From Table C.1, and more clearly from Figure 5.2 (A-C) and Figure 5.2 (D-F), it can be seen that for each scenario of hazard ratios, the required number of events and patients in the biomarker-negative subgroup is greater than in biomarker-positive subgroup.

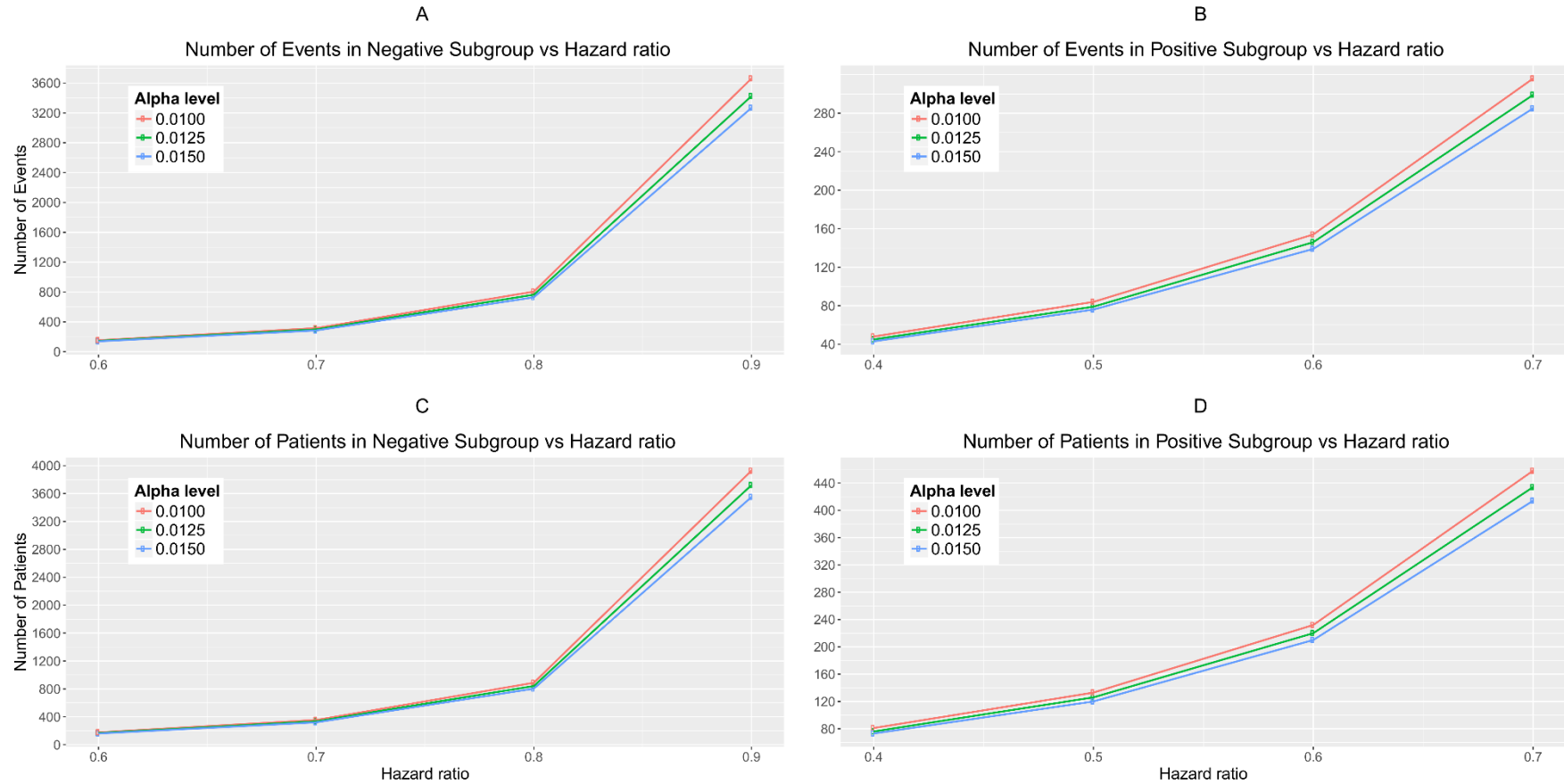


**Figure 5.2.** A, B, C represent the required number of events and D, E, F represent the required number of patients of each biomarker-defined subgroup which achieve 80% power versus the corresponding hazard ratio for each of the three scenarios of significance levels. Figure 5.2 A and D corresponds to the significance levels  $\alpha_- = \alpha_+ = 0.0125$ , Figure 5.2 B and E corresponds to the significance levels  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and Figure 5.2 C and F corresponds to the significance levels  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$ .



Figure 5.3 (A, B) represents the required number of events which achieve 80% power versus the hazard ratio for each of the three scenarios of significance levels in each biomarker-defined subgroup separately. Figure 5.3 (C, D) represents the required number of patients which achieve 80% power versus the hazard ratio for each of the three scenarios of significance levels in each biomarker-defined subgroup separately. The corresponding numerical results are presented in Table C.1.

It can be seen that for all scenarios of hazard ratios, the highest value of the number of events and patients in the biomarker-negative subgroup and in the biomarker-positive subgroup is given by  $a_- = 0.010$  and  $a_+ = 0.010$  for negative and positive patients respectively and the lowest value is given by  $a_- = 0.015$  and  $a_+ = 0.015$  for negative and positive patients respectively.

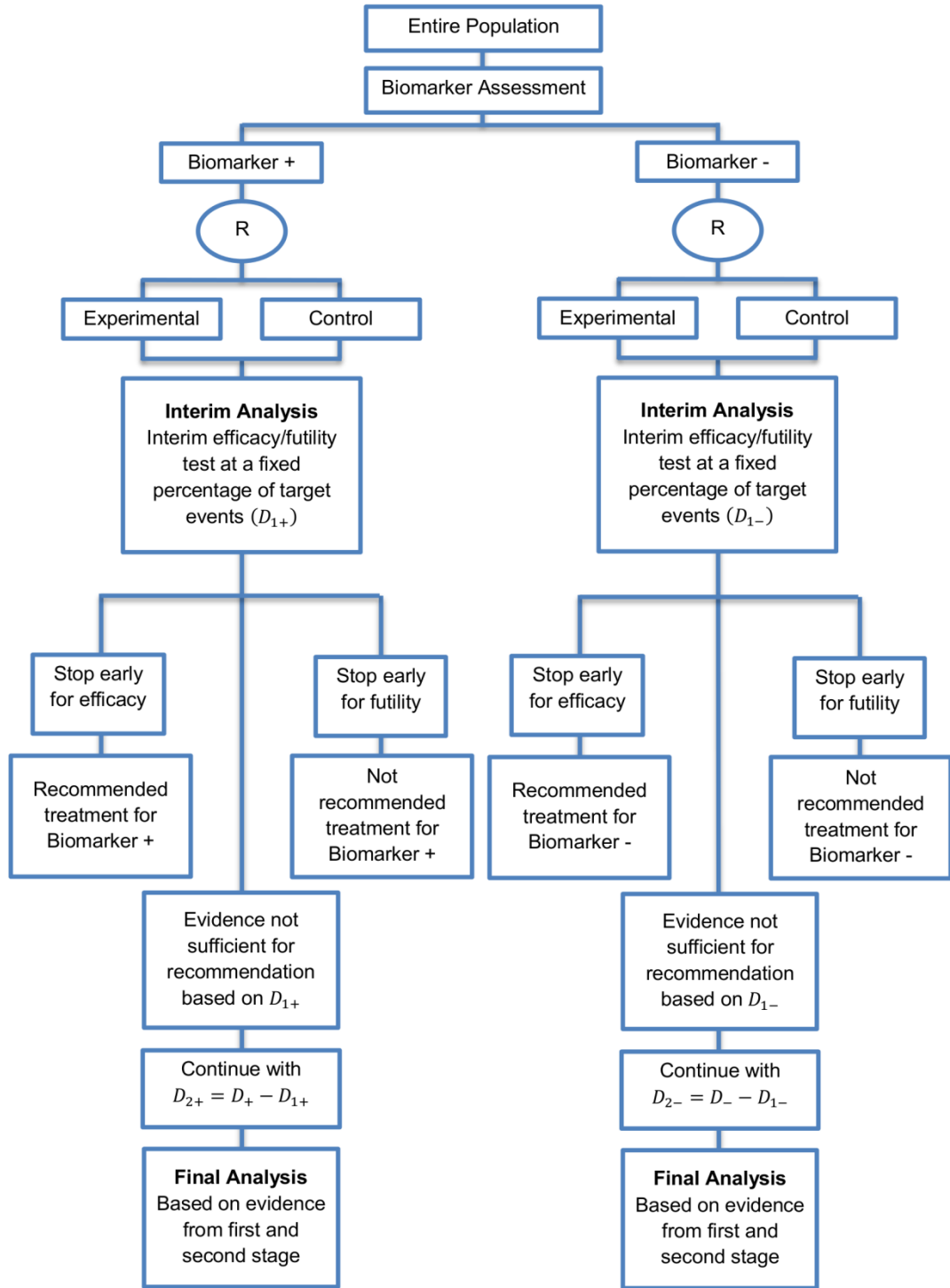


**Figure 5.3.** A, B represent the required number of events and C, D represent the required number of patients which achieve 80% power versus the hazard ratio in each of the three scenarios of significance levels for each biomarker-defined subgroup separately. Figure 5.3 A and C corresponds to the biomarker-negative subgroup and the following significance levels: (i)  $\alpha_- = 0.0125$ , (ii)  $\alpha_- = 0.015$  and (iii)  $\alpha_- = 0.010$ . Figure 5.3 B and D corresponds to the biomarker-positive subgroup and the following significance levels: (i)  $\alpha_+ = 0.0125$ , (ii)  $\alpha_+ = 0.010$  and (iii)  $\alpha_+ = 0.015$ .

### 5.2.2. An adaptive version of the Parallel Subgroup-Specific design

---

We explore a two-stage adaptive design starting with the parallel assessment of treatment effect in each biomarker-defined subgroup. In the first stage an interim analysis is included where each biomarker-defined subgroup can stop early for futility or efficacy. The interim analysis is based on a fixed and pre-specified percentage of target events. In case that we do not stop after the first stage due to early efficacy or futility, the trial continues to the second stage, testing the efficacy of the experimental treatment in each biomarker-defined subgroup separately. A graphical illustration of this strategy is given in Figure 5.4.



**Figure 5.4.** Adaptive Parallel Subgroup-Specific design. “R” refers to randomization of patients.  $D_{1+}$  and  $D_{1-}$  correspond to the target number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively at the first stage of the study.  $D_+$  and  $D_-$  correspond to the total required number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively which are planned according to the non-adaptive approach.  $D_{2+}$  and  $D_{2-}$  correspond to the number of events of the biomarker-positive subgroup and biomarker-negative subgroup respectively at the second stage of the study.

Adaptive designs differ from fixed designs in that they permit the performance of interim analyses during the course of the study leading to adaptations of hypotheses which are under investigation. Results from interim analyses are used to make a decision regarding the trial. Several sources of multiplicity problems can arise in the conduct of adaptive trial designs resulting in the inflation of the overall type I error rate (probability of a false positive result). One of the sources of type I error inflation is the adaptation of design and analysis features with combination of information across trial stages [10]. Hence, alpha-adjustment (i.e. adjustment of the alpha level at each interim analysis so that the overall type I error rate remains at the desired level) is needed so that the overall type I error rate remains under control. A variety of methods for the control of type I error rate in adaptive designs have been proposed which are thoroughly summarized by Chang (2014) [11]. Our study is based on a flexible and general approach to adaptive designs for alpha-adjustment proposed by Chang (2007) [12] in which the test statistic of the final analysis is defined as the sum of the unadjusted stagewise p-value ( $p_l$ ). More precisely, at the  $k^{\text{th}}$  stage of an adaptive design, the test statistic which can be viewed as cumulative evidence against the null is given by

$$T'_k = \sum_{l=1}^k p_l, k = 1, \dots, L =: \text{"maximum stage"}.$$

Before conducting the two-stage design, pre-specification of stopping rules and stopping boundaries for efficacy and futility are needed. Stopping probabilities (i.e. rejection probabilities), which are calculated based on the stopping boundaries, are essential operating characteristics of adaptive designs and they are classified into two types. The first type is the so-called 'efficacy stopping probability' which refers to the unconditional probability of rejecting the null hypothesis of no treatment effect, thus the trial stops in order to claim efficacy. The second type is the so-called 'futility stopping probability' which refers to the unconditional probability of not rejecting the null hypothesis of no treatment effect, thus the trial stops in order to claim futility. Hence, the following stopping boundaries should be chosen: (i) the early efficacy stopping boundaries in stage 1, i.e.  $\epsilon_{1-}$  and  $\epsilon_{1+}$  for biomarker-negative and

biomarker-positive patients respectively, (ii) the early futility stopping boundaries in stage 1, i.e.  $b_{1-}$  and  $b_{1+}$  for biomarker-negative and biomarker-positive patients respectively and (iii) the final efficacy stopping boundaries,  $\epsilon_{2-}$  and  $\epsilon_{2+}$  for biomarker-negative and biomarker-positive patients respectively.

If there is prior belief that the experimental treatment is of strong benefit to patients, then the trial should be designed without early futility stopping (i.e. we need to set a larger value for  $b_{1-}$  and/or  $b_{1+}$ ). When early efficacy stopping is allowed (e.g., to allow possibility of making treatment available to patients earlier or to allow possibility of unnecessary treatment exposure or unnecessary trial costs), then the trial should be designed with a large value of  $\epsilon_{1-}$  and/or  $\epsilon_{1+}$ .

After the appropriate choice of  $\epsilon_{1-}$ ,  $\epsilon_{1+}$  and  $b_{1-}$ ,  $b_{1+}$ , we can solve for the final efficacy stopping boundaries, i.e.  $\epsilon_{2-}$ ,  $\epsilon_{2+}$  with reference to Chang et al.'s method based on the sum of p-value. More precisely, in a clinical trial with  $k$  interim analyses, the stopping boundary can be derived by calculating the rejection probability under the null hypothesis which takes into account the stopping rules described below. The rejection probability at the  $k$ th stage is defined by  $\psi_k(\epsilon_k)$ , where

$$\begin{aligned}\psi_k(t') &= \Pr(\epsilon_1 < T'_1 < b_1, \dots, \epsilon_{k-1} < T'_{k-1} < b_{k-1}, T'_k < t') \\ &= \int_{\epsilon_1}^{b_1} \dots \int_{\epsilon_{k-1}}^{b_{k-1}} \int_{-\infty}^{t'} f_{T'_1 \dots T'_k}(t'_1, \dots, t'_k) dt'_k dt'_{k-1} \dots dt'_1\end{aligned}\quad (5.8)$$

where  $t' \geq 0$ ,  $t'_l$  ( $l = 1, \dots, k$ ) is the test statistic at the  $l$ th stage, and  $f_{T'_1 \dots T'_k}$  is the joint probability density function of  $T'_1, \dots, T'_k$ . The stopping rules for futility can be either binding or non-binding. In the non-binding rule the possibility of stopping early for futility will not be considered in the decision of the efficacy stopping boundary whereas in the binding category the futility rule is taken into account when making inference. As it is stated by Chang (2014) [11], the regulatory bodies currently adopt the non-binding futility rule in order to ensure that the familywise Type I error rate is controlled regardless of whether a decision is made to continue the trial despite a futility boundary being crossed. According to Chang (2007) [12], it is better to set

$b_{1-} = \epsilon_{2-}$  and  $b_{1+} = \epsilon_{2+}$ . Based on (5.8), according to [12] the final efficacy stopping boundaries can be found by

$$\epsilon_2 = \sqrt{(a - \epsilon_1)} + \epsilon_1,$$

where  $\epsilon_1 < a$  and  $a$  refers to the level of significance. In our case, the final efficacy stopping boundaries for each biomarker-defined subgroup with non-binding futility rule can be found by the following formulations,

$$\epsilon_{2-} = \sqrt{(a_- - \epsilon_{1-})} + \epsilon_{1-},$$

$$\epsilon_{2+} = \sqrt{(a_+ - \epsilon_{1+})} + \epsilon_{1+},$$

where  $\epsilon_{1-} < a_-$  and  $\epsilon_{1+} < a_+$ .

For the biomarker-negative subgroup of the two-stage adaptive design which tests the efficacy of the experimental treatment, the stopping rules are the following,

$$\text{Stage 1: } \begin{cases} \text{Reject the null hypothesis (stop for efficacy)} & \text{if } T'_{1-} \leq \epsilon_{1-} \\ \text{Do not reject the null hypothesis (stop for futility)} & \text{if } T'_{1-} > b_{1-} \\ \text{Continue to the second stage} & \text{if } \epsilon_{1-} < T'_{1-} \leq b_{1-} \end{cases}$$

where  $0 < \epsilon_{1-} < b_{1-} \leq 1$  and  $T'_{1-}$  refers to the test statistic as defined previously in this section in the biomarker-negative subgroup at the first stage of the study ,

$$\text{Stage 2: } \begin{cases} \text{Reject the null hypothesis (stop for efficacy)} & \text{if } T'_{2-} \leq \epsilon_{2-} \\ \text{Do not reject the null hypothesis (stop for futility)} & \text{if } T'_{2-} > \epsilon_{2-} \end{cases}$$

where  $T'_{2-}$  refers to the test statistic in the biomarker-negative subgroup at the second stage of the study.

For the biomarker-positive subgroup of the two-stage adaptive design which tests the efficacy of the experimental treatment, the stopping rules are the following,

$$\text{Stage 1: } \begin{cases} \text{Reject the null hypothesis (stop for efficacy)} & \text{if } T'_{1+} \leq \epsilon_{1+} \\ \text{Do not reject the null hypothesis (stop for futility)} & \text{if } T'_{1+} > b_{1+} \\ \text{Continue to the second stage} & \text{if } \epsilon_{1+} < T'_{1+} \leq b_{1+} \end{cases}$$

where  $0 < a_{1+} < b_{1+} \leq 1$  and  $T'_{1+}$  refers to the test statistic in the biomarker-positive subgroup at the first stage of the study,

$$\text{Stage 2: } \begin{cases} \text{Reject the null hypothesis (stop for efficacy)} & \text{if } T'_{2+} \leq \epsilon_{2+} \\ \text{Do not reject the null hypothesis (stop for futility)} & \text{if } T'_{2+} > \epsilon_{2+} \end{cases}$$

where  $T'_{2+}$  refers to the test statistic in the biomarker-positive subgroup at the second stage of the study.

We now assume the interim fraction for the biomarker-negative subgroup to be  $f_-$  which refers to a specific proportion of the required total number of events in the biomarker-negative subgroup, and the interim fraction for the biomarker-positive subgroup be  $f_+$  which refers to a specific proportion of the required total number of events in the biomarker-positive subgroup. Using these interim fractions, we calculate the target number of events for each subgroup at the interim stage (stage 1), to be:

$$\begin{aligned} D_{1-} &= D_- \times f_-, \\ D_{1+} &= D_+ \times f_+, \end{aligned}$$

for negative and positive patients respectively. The log-rank test statistics of each biomarker-defined subgroup at the first stage (interim analysis) are based on  $D_{1-}$ ,  $D_{1+}$  and given by

$$\begin{aligned} T'_{L1-} &= \sqrt{\frac{\hat{D}_{1-}}{4}} \times [\log(\hat{\theta}_-)] \sim N\left(\sqrt{\frac{D_{1-}}{4}} \times [\log(\theta_-)], 1\right), \\ T'_{L1+} &= \sqrt{\frac{\hat{D}_{1+}}{4}} \times [\log(\hat{\theta}_+)] \sim N\left(\sqrt{\frac{D_{1+}}{4}} \times [\log(\theta_+)], 1\right), \end{aligned}$$

for the biomarker-negative subgroup and biomarker-positive subgroups respectively. One sided p-value corresponding to the observed values  $t'_{L1-}$  and  $t'_{L1+}$  of the test statistics of each biomarker-defined subgroup in stage 1 are given by



$$p_{1-} = Pr(T'_{L1-} \geq t'_{L1-} | H_{0,biom-}),$$

$$p_{1+} = Pr(T'_{L1+} \geq t'_{L1+} | H_{0,biom+}),$$

for the biomarker-negative subgroup and biomarker-positive subgroups respectively.

In the first interim analysis, the test statistic is equal to the p-value at stage 1; hence in our simulation study we proceed with the following rules: If  $p_{1-} > b_{1-}$  or/and  $p_{1+} > b_{1+}$  then the study which is testing the efficacy of the experimental treatment in biomarker-negative subgroup and/or biomarker-positive subgroup is stopped for futility at stage 1. If  $p_{1-} \leq \epsilon_{1-}$  and/or  $p_{1+} \leq \epsilon_{1+}$  then the study which is testing the efficacy of the experimental treatment in biomarker-negative subgroup and/or biomarker-positive subgroup is stopped for efficacy at stage 1. Otherwise, if  $\epsilon_{1-} < p_{1-} \leq b_{1-}$  and/or if  $\epsilon_{1+} < p_{1+} \leq b_{1+}$ , the study which is testing the treatment effect in each biomarker-defined subgroup continues to the second stage.

The log-rank test statistics of each biomarker-defined subgroup at the second stage of the study are given by

$$T''_{L2-} = \sqrt{\frac{(\hat{D}_- - \hat{D}_{1-})}{4}} \times [\log(\hat{\theta}_-)] \sim N\left(\sqrt{\frac{(D_- - D_{1-})}{4}} \times [\log(\theta_-)], 1\right),$$

$$T''_{L2+} = \sqrt{\frac{(\hat{D}_+ - \hat{D}_{1+})}{4}} \times [\log(\hat{\theta}_+)] \sim N\left(\sqrt{\frac{(D_+ - D_{1+})}{4}} \times [\log(\theta_+)], 1\right),$$

for the biomarker-negative subgroup and biomarker-positive subgroups respectively. One-sided p-values corresponding to the observed values  $t''_{L2-}$  and  $t''_{L2+}$  of the test statistics of each biomarker-defined subgroup in stage 2 are given by

$$p_{2-} = Pr(T''_{L2-} \geq t''_{L2-} | H_{0,biom-}),$$

$$p_{2+} = Pr(T''_{L2+} \geq t''_{L2+} | H_{0,biom+}),$$

for the biomarker-negative subgroup and biomarker-positive subgroup respectively. The test statistic of the final analysis for each biomarker-defined subgroup is based on the sum of stagewise p-value and can be given by

$$T'_{-} = p_{1-} + p_{2-},$$

$$T'_{+} = p_{1+} + p_{2+},$$

for the biomarker-negative subgroup and biomarker-positive subgroup respectively.

#### 5.2.2.1. Simulation Study 2

---

To investigate the effect of introducing an adaptive element to our study design, we have conducted a second simulation study which is performed by using the R statistical software. To do this we assume the same total number of events and patients as we did for Simulation Study 1 where our design was not adaptive, therefore making the same assumptions regarding significance levels for biomarker-negative and biomarker-positive subgroups ( $a_{-}, a_{+}$ ), hazard ratios ( $\theta_{-}, \theta_{+}$ ), median survival time in Control group ( $m_{c-}, m_{c+}$ ), accrual period ( $T_{-}, T_{+}$ ) and follow-up period ( $\tau_{-}, \tau_{+}$ ) as we did previously. Therefore we assume accrual time ( $T_j$ ) to be 18 months for both subgroups, follow-up time ( $\tau_j$ ) to be 12 months for both subgroups and consider the following four different scenarios for the hazard ratios, (i)  $\theta_{-} = 0.6$  and  $\theta_{+} = 0.4$ , (ii)  $\theta_{-} = 0.7$  and  $\theta_{+} = 0.5$ , (iii)  $\theta_{-} = 0.8$  and  $\theta_{+} = 0.6$  and (iv)  $\theta_{-} = 0.9$  and  $\theta_{+} = 0.7$ . For each scenario of hazard ratios we again assume three different scenarios for significance levels for the biomarker-negative and biomarker-positive subgroups, i.e. (i)  $a_{-} = a_{+} = 0.0125$ , (ii)  $a_{-} = 0.015$  and  $a_{+} = 0.010$  and (iii)  $a_{-} = 0.010$  and  $a_{+} = 0.015$ . For each hazard ratios and significance level combination explored previously, we test the implication of different percentages of the information fraction. The different information fractions considered are as follows: (i)  $f_{-} = f_{+} = 25\%$ , (ii)  $f_{-} = f_{+} = 50\%$  and (iii)  $f_{-} = f_{+} = 75\%$ . Our aim is to explore the impact of these different information fractions on study power as well as on the stopping probabilities for futility ( $FSP_j$ ) and efficacy ( $ESP_j$ ).

In our simulation study for all the scenarios of hazard ratios for each biomarker-defined subgroup we used a high value of early efficacy stopping boundaries, i.e.  $\epsilon_{1+}, \epsilon_{1-}$  and thus a high value of early futility stopping boundaries, i.e.  $b_{1+}, b_{1-}$  as it is believed that the experimental treatment is promising. Thus, for the three cases of significance levels for each biomarker-defined subgroup we have set the following stopping boundaries:

(i) when  $a_- = a_+ = 0.0125$ ,

for  $\epsilon_{1+} = 0.0080$  we get  $b_{1+} = \epsilon_{2+} = \sqrt{(a_+ - \epsilon_{1+})} + \epsilon_{1+} = 0.1029$ ,

for  $\epsilon_{1-} = 0.0070$  we get  $b_{1-} = \epsilon_{2-} = \sqrt{(a_- - \epsilon_{1-})} + \epsilon_{1-} = 0.1129$ ,

(ii) when  $a_- = 0.015$  and  $a_+ = 0.010$ ,

for  $\epsilon_{1+} = 0.0080$  we get  $b_{1+} = \epsilon_{2+} = \sqrt{(a_+ - \epsilon_{1+})} + \epsilon_{1+} = 0.0527$ ,

for  $\epsilon_{1-} = 0.0070$  we get  $b_{1-} = \epsilon_{2-} = \sqrt{(a_- - \epsilon_{1-})} + \epsilon_{1-} = 0.0964$ ,

(iii) when  $a_- = 0.010$  and  $a_+ = 0.015$ ,

for  $\epsilon_{1+} = 0.0080$  we get  $b_{1+} = \epsilon_{2+} = \sqrt{(a_+ - \epsilon_{1+})} + \epsilon_{1+} = 0.0917$ ,

for  $\epsilon_{1-} = 0.0070$  we get  $b_{1-} = \epsilon_{2-} = \sqrt{(a_- - \epsilon_{1-})} + \epsilon_{1-} = 0.0618$ .

In all cases we have used a slightly lower value for  $\epsilon_{1-}$  (i.e. 0.007) assuming that it is believed that the experimental treatment is less promising in biomarker-negative subgroup as compared to the biomarker-positive subgroup.

Different values of stopping boundaries could be used for each assumed scenario of hazard ratios and significance levels based on how promising the experimental treatment seems to be in each subgroup. However, for simplicity, in our study we set only one value of  $\epsilon_{1+}$  for biomarker-positive subgroup and only one value of  $\epsilon_{1-}$  for biomarker-negative subgroup in all cases of hazard ratios.

The general efficiency related to the cost and time of the trial can be seen from the expected number of events and expected sample size of the trial which are calculated by using the futility and efficacy stopping probabilities. The expected sample size is defined by Chang (2014) [11] as a function of the effect size and its uncertainty, which are unknown. Hence, apart from the stopping probabilities and power, this simulation study also provides the average expected number of events for each biomarker-defined subgroup ( $D_-^{exp}, D_+^{exp}$ ). Based on the futility ( $FSP_j$ ) and efficacy stopping probabilities ( $ESP_j$ ), we also calculate the average expected sample size for each biomarker-defined subgroup, i.e.

$$N_-^{exp} = [(ESP_- + FSP_-) \times N_{1-}] + \{[1 - (ESP_- + FSP_-)] \times N_-\},$$

$$N_+^{exp} = [(ESP_+ + FSP_+) \times N_{1+}] + \{[1 - (ESP_+ + FSP_+)] \times N_+\},$$

for negative and positive patients respectively. Assuming that we have constant accrual, we can calculate the expected duration of the trial for testing the treatment effect in each biomarker-defined subgroup, i.e.

$$T_-^{exp} = [(ESP_- + FSP_-) \times (T_- + \tau_-) \times f_-] + \{[1 - (ESP_- + FSP_-)] \times (T_- + \tau_-)\},$$

$$T_+^{exp} = [(ESP_+ + FSP_+) \times (T_+ + \tau_+) \times f_+] + \{[1 - (ESP_+ + FSP_+)] \times (T_+ + \tau_+)\},$$

for negative and positive patients respectively.

#### 5.2.2.2. Results from simulation study 2

---

Table 5.1 in this section and Tables C.2-C.4 in Appendix C provide the simulation results drawn from 10000 iterations for each scenario of hazard ratios and significance levels, for each different assumed percentage of information fraction, i.e. (i)  $f_- = f_+ = 25\%$ , (ii)  $f_- = f_+ = 50\%$  and (iii)  $f_- = f_+ = 75\%$ . We report the expected number of events and patients, expected total study duration, futility stopping probability, efficacy stopping probability and total power of the study in addition to the required number of patients and events (as presented in Table C.1 for the fixed Parallel Subgroup-Specific design).

**Table 5.1.** Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 1 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the first scenario of hazard ratios and significance levels are also presented.

		Simulation setting		Number		Simulated Power					
Group of patients	Significance level	Hazard ratio	Required	Required	Expected Total	Expected	Expected	FSP	ESP	Power	
			Number of events	Number of patients	study period (months)	Number of events	Number of patients				
25%	Biomarker-negative	0.0125	0.6	146	168	17.6	86	99	0.3694	0.1810	0.5659
	Biomarker-positive	0.0125	0.4	45	76	16.9	25	43	0.3947	0.1894	0.5371
	Entire population	0.025	-	191	244	-	111	142	-	-	-
	Biomarker-negative	0.015	0.6	139	160	16.8	78	90	0.4172	0.1692	0.4999
	Biomarker-positive	0.010	0.4	48	81	14.0	22	38	0.5071	0.2059	0.4257
	Entire population	0.025	-	187	241	-	100	128	-	-	-
	Biomarker-negative	0.010	0.6	154	177	14.9	77	88	0.4821	0.1886	0.4454

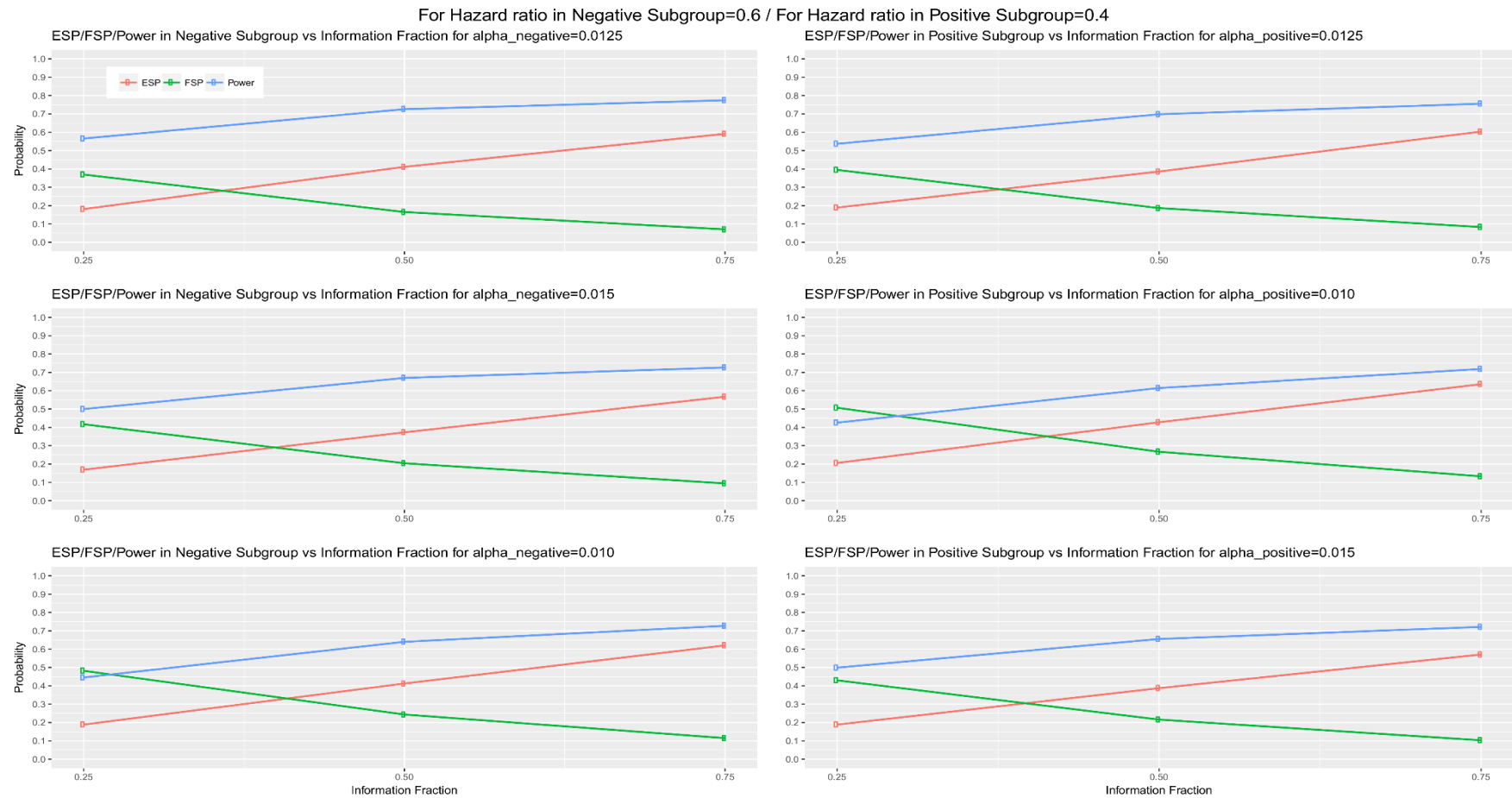
	Biomarker-positive	0.015	0.4	43	72	16.1	23	39	0.4300	0.1885	0.4990
	Entire population	0.025	-	197	249	-	100	127	-	-	-
50%	Biomarker-negative	0.0125	0.6	146	168	21.4	106	120	0.1650	0.4114	0.7259
	Biomarker-positive	0.0125	0.4	45	76	21.4	32	54	0.1865	0.3859	0.6982
	Entire population	0.025	-	191	244	-	138	138	-	-	-
	Biomarker-negative	0.015	0.6	139	160	21.3	99	114	0.2044	0.3730	0.6697
	Biomarker-positive	0.010	0.4	48	81	19.6	31	53	0.2670	0.4279	0.6146
	Entire population	0.025	-	187	241	-	130	167	-	-	-
	Biomarker-negative	0.010	0.6	154	177	20.2	103	119	0.2435	0.4126	0.6400
	Biomarker-positive	0.015	0.4	43	72	21.0	30	50	0.2159	0.3873	0.6555
	Entire population	0.025	-	197	249	-	133	169	-	-	-

75%	Biomarker-negative	0.0125	0.6	146	168	25.0	122	140	0.0704	0.5915	0.7743
	Biomarker-positive	0.0125	0.4	45	76	24.9	37	63	0.0830	0.6036	0.7558
	Entire population	0.025	-	191	244	-	159	203	-	-	-
	Biomarker-negative	0.015	0.6	139	160	25.0	116	134	0.0943	0.5674	0.7266
	Biomarker-positive	0.010	0.4	48	81	24.2	39	65	0.1330	0.6356	0.7185
	Entire population	0.025	-	187	241	-	155	199	-	-	-
	Biomarker-negative	0.010	0.6	154	177	24.5	126	145	0.1148	0.6204	0.7273
	Biomarker-positive	0.015	0.4	43	72	24.9	36	60	0.1030	0.5708	0.7214
	Entire population	0.025	-	197	249	-	162	205	-	-	-

From Table 5.1 and Tables C.2-C.4, it can be seen that the futility stopping probability of each biomarker-defined subgroup at each significance level decreases when the percentage of information fraction increases. In contrast, the efficacy stopping probability, the total power of the study, the sample size and the number of events increase with the increase of the information fraction.

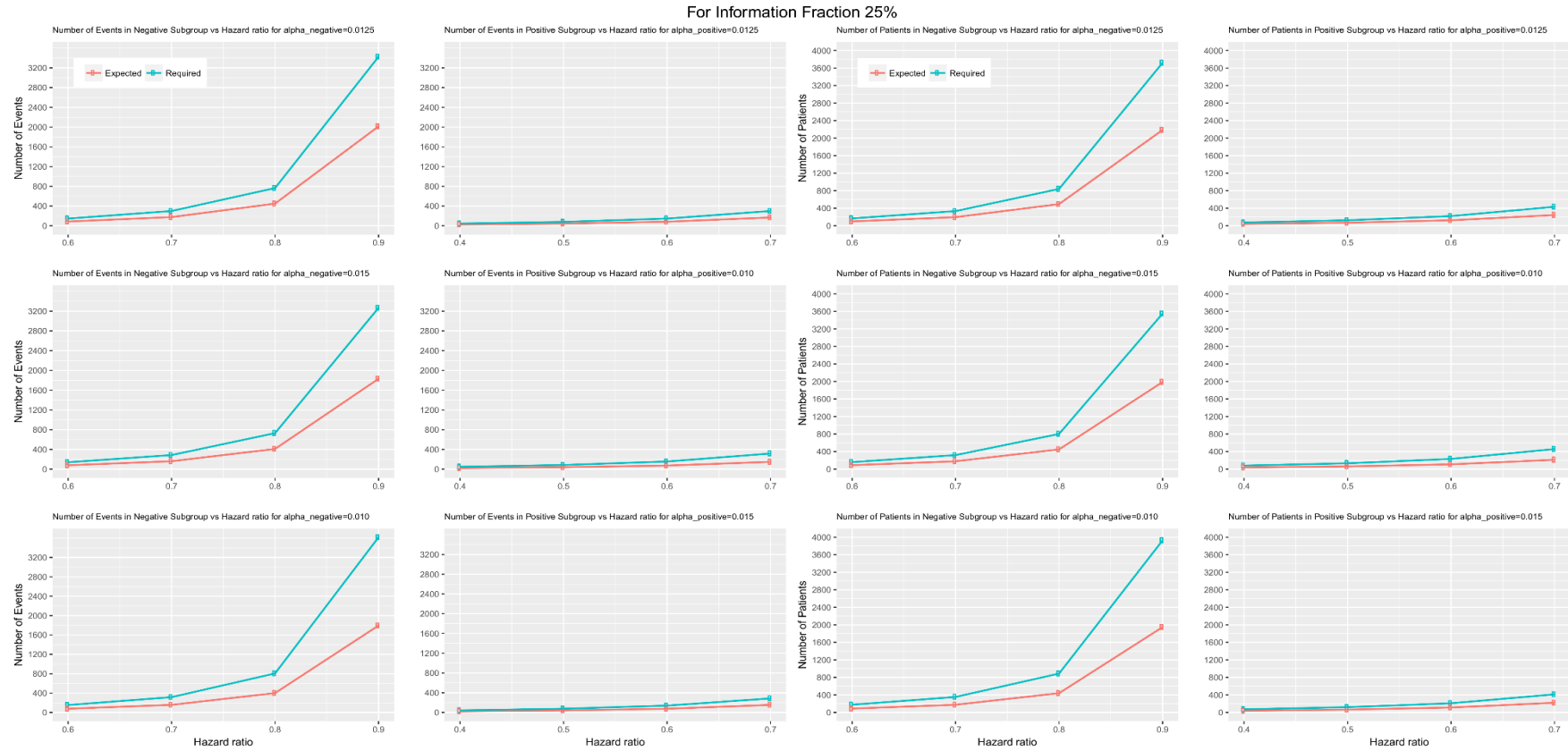
When the interim fraction is set to 25% of the required total number of events, the simulation results indicate that the trial is underpowered. When the interim fraction is based on 50% of the required total number of events we still do not have a gain in power compared to the nominal level 80%, however, it achieves approximately 73% and 70% power in biomarker-negative and biomarker-positive subgroup respectively in all scenarios of hazard ratios when  $a_- = a_+ = 0.0125$ . Compared to the 30 months overall duration of the fixed design, in the above scenario the expected overall duration is reduced approximately by one third. When the interim fraction is based on 75% of the required total number of events, we can achieve higher level of power, i.e. approximately 77% and 76% power in biomarker-negative and biomarker-positive subgroup respectively in all scenarios of hazard ratios when  $a_- = a_+ = 0.0125$ . Compared to the 30 months overall duration of the fixed design, in the above scenario the expected overall duration does not exceed 25 months. These results for the first scenario of hazard ratios are graphically represented in Figure 5.5. The results for the remaining scenarios of hazard ratios are graphically represented in Figures C.1-C.3 in Appendix C.





**Figure 5.5.** Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 1 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively.

Figure 5.6 shows the expected number of events and patients of the adaptive Parallel Subgroup-Specific design and the required number of events and patients of the fixed Parallel Subgroup-Specific design for each biomarker-defined subgroup versus the corresponding hazard ratios for the first level of information fraction (i.e. 25%). Figures C.4-C.5 provided in Appendix C show the expected number of events and patients of the adaptive Parallel Subgroup-Specific design and the required number of events and patients of the fixed Parallel Subgroup-Specific design for each biomarker-defined subgroup versus the corresponding hazard ratios for the second and third level of information fraction (i.e. 50%, 75%). Figure 5.6 and Figures C.4-C.5 show that the expected number of events for both biomarker-defined subgroups in all cases of hazard ratios is lower than the required number of events. Additionally, in all cases of interim fraction and significance levels, both the required and the expected number of events are greater in biomarker-negative subgroup as compared to the biomarker-positive subgroup. Furthermore, they show that the expected number of patients for both biomarker-defined subgroups in all cases of hazard ratios is lower than the required number of patients. In addition, in all cases of interim fraction and significance levels, both the required and the expected number of patients are greater in biomarker-negative subgroup as compared to the biomarker-positive subgroup.



**Figure 5.6.** Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 25%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively.

### 5.3. Discussion

---

To conclude, in the current chapter we have considered a fixed design which evaluates the efficacy of treatment in each biomarker-defined subgroup and an adaptive approach which involves early stopping of the trial due to efficacy or futility. The scope of the simulation study of the fixed design is to investigate the power under different scenarios in a given simulation setting which takes into account accrual and follow-up of patients. Next, we extend the fixed design into a two-stage design with interim analysis used for decision making. The aim of the two-stage version was to investigate the general efficiency of the study by calculating the expected number of events and patients as well as stopping probabilities, overall power and expected duration of the study under different scenarios of the information fraction (i.e. specific proportion of the required total number of events applied in interim analysis). The tests between the two biomarker-defined subgroups are independent and one could present the results for only one of them. However, this could have added confusion regarding how control of the overall alpha is handled. Additionally, presenting results for both subgroups adds completeness to our study. It would be worth mentioning that early stopping in one biomarker defined-subgroup and not the other will still require screening patients from the entire population. Hence, this should be taken into account when conducting a real clinical trial.

We programmed the simulation studies in R statistical software. Our results indicate that when the information fraction used in interim analysis is low (25%) the study will not achieve adequate power. However, for equally allocated significance levels we can achieve sufficient power (between 70% and 80%) if the specific proportion of the required total number of events is equal or greater than 50%.

One significant challenge encountered when conducting such flexible trial designs is the multiplicity issues which should be carefully considered, e.g., in our simulation study we took into account the control of type I error not only because we had to combine information from both stages of the design but because we tested

more than one subgroup of interest at the same time. Several methods have been proposed recently for multiplicity adjustment and suggestions of appropriate stopping boundaries when an interim analysis is introduced in the study design. Thus, the implications of the operating characteristics to the decision-making when these methods are applied should be explored to get the optimal results. Our simulation studies are limited to particular methods, however, they give us a general insight into the implications of an event driven design for which the decision making is based on the results of an interim analysis. We explore a fixed versus an adaptive approach in a popular biomarker-guided clinical trial setting which to our best knowledge has not been investigated yet.

The adaptive version of the parallel subgroup-specific design could be extended by using a blinded sample size re-estimation approach [11] that reestimated based on the event rate. More precisely, the idea behind this method is that at the interim analysis -which in our case will be based on a fixed percentage of target events- we can allow for re-evaluation of the sample size when there is uncertainty about the event rate..

Knowledge on how to design, implement and analyse biomarker-guided clinical trials is essential for testing the effectiveness of a biomarker-guided approach to treatment. The proper choice and use of such designs can increase the probability of success of clinical trials resulting in the development of novel treatments. Adaptive designs might be more complex and need more time during the planning process due to several simulations of possible scenarios that should be conducted aiming to investigate the statistical properties of the design under specific situations. However, they will continue to be an attractive approach of clinical development as they can lead to potential reduction in cost and time compared to a non-adaptive approach.

In the next chapter (Chapter 6) we will show how a researcher could choose the most appropriate design among several non-adaptive trials. A series of strategies have been developed that could be used in future for the decision making.

## 5.4. References

---

1. George SL. Statistical issues in translational cancer research. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(19):5954-8. doi: 10.1158/1078-0432.CCR-07-4537.
2. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Biomarker-Guided Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *PLoS ONE*. 2016; 11(2):e0149803. doi: 10.1371/journal.pone.0149803.
3. Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *Journal of Personalized Medicine*. 2017; 7(1). doi: 10.3390/jpm7010001.
4. Freidlin B, Sun Z, Gray R, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2013; 31(25):3158-61. doi: 10.1200/JCO.2012.48.3826.
5. Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature reviews Clinical oncology*. 2014; 11(2):81-90. doi: 10.1038/nrclinonc.2013.218.
6. Freidlin B, Korn EL, Gray R. Marker Sequential Test (MaST) design. *Clinical trials (London, England)*. 2014; 11(1):19-27. doi: 10.1177/1740774513503739.
7. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal of biopharmaceutical statistics*. 2009; 19(3):530-42. doi: 10.1080/10543400902802458.
8. Collette L. Modelling survival data in medical research. 2nd ed. Boca Raton, Fla: Chapman & Hall/CRC; 2003.

9. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 3rd ed. New York, NY: Springer; 2012.
10. Wassmer G, Brannath W. Group sequential and confirmatory adaptive designs in clinical trials. Switzerland: Springer; 2016.
11. Chang M. Adaptive Design Theory and Implementation Using SAS and R, Second Edition. 2nd ed. London: CRC Press; 2014.
12. Chang M. Adaptive design method based on sum of p-values. STATISTICS IN MEDICINE. 2007; 26(14).

## Chapter 6. Deciding on the best biomarker-guided trial design: a case study

---

### 6.1. Introduction

---

In the previous chapter we explored statistical properties of a fixed and adaptive design through a particular simulation setting. In the current chapter we reconsider the most appropriate design for a clinical trial previously proposed for testing whether a genotype-guided treatment strategy results in reduced rate of relapse in alcohol-dependent patients. The design originally proposed was not feasible due to low prevalence of the genetic marker. Here, we explore the various biomarker-guided non-adaptive trial designs and apply a strategy to choose the most appropriate for the trial in question. To our best knowledge, there is no suggested strategy in literature to help in decision making. We also apply statistical techniques to calculate the necessary sample size, considering both binary and time-to-event outcomes. For time-to-event outcomes we propose an approximation for the calculation of the required number of events. In addition, an adaptive version of the proposed design is explored in Chapter 7, to address a degree of uncertainty that surrounds the assumed effect size.

### 6.2. Background to the proposed trial (STRONG trial)

---

Existing research: Alcohol misuse accounts for almost 10% of the total UK disease burden [1], with an estimated 22,000 deaths each year [2]. Over 12% of Accident and Emergency attendances are attributed to alcohol, increasing to 70% at peak times [3]. The annual cost of alcohol-related harm (e.g. violent incidents, anti-social behaviour, working days lost, hospital admissions, domestic violence, impact on the family expenditure on specialist alcohol treatment, premature deaths) to the NHS has been estimated to be £2 billion [2]. An assessment of alcohol-related harm showed that 38% of men and 16% of women (aged 16-64 years) have an alcohol use disorder, equating to approximately 8.2 million people in England. Of most



concern is that an estimated 3-5% of the population of England (1.1 million people) are alcohol dependent.

**Pharmacotherapy of alcohol dependence:** Clinical trials have shown both acamprosate and naltrexone to be superior to placebo in reducing the rate of relapse to drinking following detoxification [4, 5]. Importantly, both naltrexone and acamprosate given to patients in the presence of supportive interventions are effective in treating patients who might otherwise not receive treatment [6]. NICE guidance (CG115) [6] recommends therapy with either acamprosate or naltrexone in conjunction with psychological support. However, while acamprosate is licensed in the UK, naltrexone is not. Even though a US Randomized Controlled Trial [7] (COMBINE: Combined Pharmacotherapies and Behavioural Interventions for Alcohol Dependence Study) reported naltrexone either alone or combined with behavioural intervention to be superior to acamprosate and placebo (Naltrexone reduced hazard of time until a heavy drinking day (hazard ratio, 0.72; 97.5% CI, 0.53 - 0.98; p-value=0.02), most evident in those receiving medical management but not combined behavioural intervention (CBI) but that acamprosate had no significant effect on drinking vs placebo (either by itself or with any combination of naltrexone, CBI, or both), recent meta-analyses [4, 5] have shown both naltrexone and acamprosate to be effective and safe strategies in alcoholism treatment. More precisely, the first meta-analysis [5] which was based on a total of 50 RCTs with 7793 participants has shown that naltrexone reduced the risk of heavy drinking compared to placebo (relative risk, 0.83; 95% CI, 0.76 -0.90). The second meta-analysis [4] which was based on a total of 24 RCTs with 6915 participants has shown that acamprosate reduced the risk of any drinking compared to placebo (relative risk, 0.86; 95% CI, 0.81-0.91). However, adjunct therapy is rarely offered in hospital settings as indicated by NICE (CG115). Given the potential benefits of the use of pharmacotherapy in conjunction with psychosocial support in treating alcohol dependence, this represents an important area where better evidence would improve management of the condition in acute hospitals.

**Variability in treatment response to pharmacotherapy of alcohol dependence:** Some studies have highlighted that the effect size for response to naltrexone and acamprosate over placebo is in the small-to-moderate range [8, 9]. More specifically, in the systematic review by Bouza et al. (2004) in which 33 studies were included, "Acamprosate was associated with a significant improvement in abstinence rate (odds ratio, 1.88; 95% CI, 1.57 to 2.25; p-value<0.001). Short-term administration of naltrexone reduced the relapse rate significantly (odds ratio, 0.62; 95% CI, 0.52 to 0.75; p-value<0.001), but was not associated with a significant modification in the abstinence rate (odds ratio, 1.26; 95% CI, 0.97 to 1.64; p-value=0.08). There were insufficient data to ascertain naltrexone's efficacy over more prolonged periods. Acamprosate had a good safety pattern and was associated with a significant improvement in treatment compliance (odds ratio, 1.29; 95% CI, 1.13 to 1.47; p-value<0.001). Naltrexone's side effects were more numerous, yet the drug was nevertheless tolerated acceptably without being associated with a lower adherence to treatment (odds ratio, 0.94; 95% CI, 0.80 to 1.1; p-value=0.5). However, overall compliance was relatively low with both medications" [8]. In Srisurapanont et al. [9] in which a total of 2861 patients in 24 RCTs presented in 32 papers were included, it is stated that "For short-term treatment, naltrexone significantly decreased relapses (relative risk, 0.64; 95% CI, 0.51–0.82), but not return to drinking (relative risk, 0.91; 95% CI, 0.81–1.02). Short-term treatment of naltrexone significantly increased nausea, dizziness, and fatigue in comparison to placebo (relative risks, 95% CI: 2.14, 1.61 to 2.83; 2.09, 1.28 to 3.39; 1.35, 1.04 to 1.75). Naltrexone administration did not significantly diminish short-term discontinuation of treatment (relative risk, 0.85; 95% CI, 0.70–1.01)". Studies have also shown that certain subpopulations respond better to pharmacotherapy than others [10]. Indeed, these findings point to the need for better strategies to stratify patients so that those most likely to respond to alcohol pharmacotherapy can be identified.

**Predictors of naltrexone response:** It is clear that not all individuals with alcohol dependence respond to naltrexone in a similar manner. Various factors have been suggested to be important predictors of naltrexone response; these include clinical factors such as family history of alcoholism (odds ratio, 2.084; 95% CI, 1.189

to 3.653;  $p$ -value= 0.010) [10], early age at onset of drinking problems (odds ratio, 2.004; 95% CI, 1.150 to 3.491;  $p$  – value =0.014) [10], comorbid use of other drugs of abuse (odds ratio, 6.348; 95% CI, 2.159 to 18.668;  $p$ -value<0.001) [10], and importantly, genetic factors such as mu-opioid receptor polymorphisms (odds ratio, 5.75; 95% CI, 1.88 to 17.54 [11], odds ratio, 6.28; 95% CI, 1.94 to 20.34 [12], odds ratio, 0.77; 95% CI, 0.28 to 2.15 [13]).

One of the most consistent predictors of naltrexone response is a missense single nucleotide polymorphism (SNP) in the *OPRM1* gene, which encodes the mu-opioid receptor type 1, the site of action of naltrexone [13]. This SNP results in the substitution of asparagine to aspartate at position 40 (Asn40Asp; rs1799971; A to G) resulting in structural variation in the receptor's extracellular domain. Asn40Asp is defined as a functional polymorphism of the mu-opioid receptor gene (*OPRM1*) and Asp40 refers to the mutant allele. Alcohol is known to increase the release of endogenous opioids such as  $\beta$ -endorphin and enkephalin in humans; blockade of opioid receptors with naltrexone leads to less alcohol-induced pleasure and intoxication, and ultimately, less craving and relapse [14].

Ray and Hutchison [15, 16] in two separate lab-based placebo-controlled experiments in humans reported that the effect of naltrexone on blunting alcohol induced highs was stronger in those who carried the Asp40 allele. A similar though distinct SNP which similarly affects the ability of the receptor to bind the endogenous ligand  $\beta$ -endorphin has also been reported to reduce alcohol-induced stimulation and alcohol consumption in rhesus monkeys [17]. In 2003, Oslin et al. [13], as part of a sub-study to a clinical trial reported that treatment-seeking alcohol-dependent individuals (African Americans and European Americans) with at least one copy of the Asp40 allele responded to naltrexone better than those without the allele [13]. More precisely, subjects who did not relapse compared to those who were homozygous for the Asn40 allele (odds ratio, 3.52; 95% CI interval, 1.03–11.96), and time to first relapse in the naltrexone-treated subjects was significantly longer in those with the Asp40 variant (hazard ratio, 2.79; 95% CI, 1.05-7.41). However, a subset analysis of the Veterans Affairs Naltrexone Cooperative Study [18], which included

215 alcohol-dependent male subjects, did not find any association between the carriage of Asp40 allele and rate of return to heavy drinking following naltrexone treatment (SNP x treatment effect  $p\text{-value} \geq 0.05$ ). Although the available patient sample was larger than that in the study by Oslin et al. [13], the authors highlighted limited statistical power for detecting a SNP x treatment interaction effect. Apart from the limited statistical power, the authors reported also differences in clinical makeup of the sample between their study and that by Oslin et al., in particular a higher proportion of participants with severe alcohol dependency, as possible reasons for the failure to find an association. Anton et al. performed a retrospective pharmacogenetic analysis [11] of the COMBINE cohort ( $n=604$ ; Caucasians only), one of the largest trials addressing the effectiveness of naltrexone against acamprosate and behavioural intervention [7]. This study found that individuals who carried at least one Asp40 mutant allele showed the best clinical outcome in the naltrexone treated arm (87.1% good outcome in Asp40 mutant allele carriers vs 54.8% good outcome in non-carriers; odds ratio, 5.75; 95% CI, 1.88-17.54). In the study, “good outcome” was defined as “abstinent or moderate drinking without problems, a maximum of 11 (women) or 14 (men) drinks per week, with no more than 2 days on which more than 3 drinks (women) or 4 drinks (men) were consumed, and 3 or fewer alcohol-related problems endorsed on the Drinker Inventory of Consequences scale during the last 8 weeks of treatment”. The subset of individuals with at least one copy of the Asp40 allele who were treated with naltrexone also had a reduced number of heavy drinking days and an increased number of abstinent days over time than did Asp40 carriers who were treated with placebo and Asp40 non-carriers (Asn40 homozygotes) who were treated with either naltrexone or placebo. A further haplotype-based sub-study reported similar results but also observed that the Asn40Asp SNP contributed most to the haplotype function [12]. The Asn40Asp SNP is also important in predicting treatment response in other ethnicities. A Korean study reported significantly lower rates of relapse (odds ratio of Asn40 patients versus Asp40 variant patients, 10.608;  $p\text{-value}=0.072$ ) to drinking in individuals who carry the Asp40 mutant allele [19], whilst a recent double blind, randomized, placebo-controlled laboratory trial of naltrexone in non-treatment-seeking Asian American

heavy drinkers found Asp40 mutant allele carriers to experience lower alcohol craving on naltrexone than non-carriers which was indicated by the statistically significant medication  $\times$  genotype interaction ( $p$ -value $<0.05$ ) [20].

With 3 out of 4 retrospective clinical sub-studies [11, 13, 18, 19] and several laboratory based studies [15, 16, 20] supporting a role for Asn40Asp SNP in naltrexone response, there is accumulating evidence, and biological plausibility, of the importance of this polymorphism in predicting naltrexone response in alcohol dependent patients. However, in order to demonstrate its clinical utility in guiding treatment, there is a need for an appropriately powered RCT designed to evaluate stratification of patients. To our knowledge, there is only one prospective RCT currently ongoing in the US (Clinical Trials.Gov Identifier: NCT00831272) where genotype-based stratification of naltrexone treatment in alcohol dependent patients is being compared to placebo in 340 patients. Given the 20% frequency of Asp40 allele in European Caucasians, it is thought that a prospective RCT with adequate power to detect clinically meaningful differences in the outcomes is required.

**Predictors of acamprosate response:** Factors suggested to be important in acamprosate response include a typological differentiation of chronic alcoholism [21, 22] (those with a more severe dependence and greater withdrawal syndrome [type I]; and those with anxiety [type II], and lesser prevalence of baseline somatic distress [21]. In an analysis by Kiefer et al. [21], out of a total of 143 participants, 101 subjects were typologized as type I and 42 subjects as type II based on Cloninger's criteria for alcoholism typology [21]. In the placebo treatment group a significant association was found between type and time to relapse as well as time to first drink. The same association was not found, however, for patients in the naltrexone, acamprosate and naltrexone plus acamprosate treatment groups. In addition, for type II participants, naltrexone, acamprosate and naltrexone plus acamprosate treatments were more effective in terms of time to first drink and time to relapse as compared to the placebo treatment group. Genetic predictors of acamprosate response are not well reported; however, evidence for the role of genetic factors is emerging. A 2011 study reported an intronic SNP in the GATA binding protein 4 gene (*GATA4*) to be associated with

relapse in acamprosate treated alcohol dependent patients (odds ratio, 2.255; 95% CI, 1.385–3.670) [23]. In the aforementioned study, 374 patients were included in the investigation and they were all participants of the PREDICT (patients with renal impairment and diabetes undergoing computed tomography) study for whom genotype information was available. The PREDICT study (Identifier: NCT00289614) was conducted at 23 sites in North America and China. GATA4 is a transcription factor regulating the transcription of atrial natriuretic peptide which is known to be involved in the pathophysiology of alcohol dependence. In addition, variation in genes involved in glutamatergic and GABAergic pathways through which acamprosate moderates neurochemical changes may also contribute to variability in its response, according to a Dutch study [24]. However, it is important to note that (a) studies undertaken so far have been relatively small; and (b) there has been no independent replication. There is thus a need for further evidence to support the role of genetic factors in predicting response to acamprosate.

Furthermore, given that both naltrexone and acamprosate appear similar in terms of therapeutic efficacy in unstratified patients with alcohol dependency, a design has been chosen that allows a comparison with acamprosate to gauge the benefit of stratification, maintain equipoise, and lay the foundations for future work which will involve other markers (clinical and laboratory) to further stratify, and hence optimize, treatment in this difficult-to-treat patient group.

#### 6.2.1. Previously proposed randomized controlled trial of a stratified approach to Naltrexone treatment

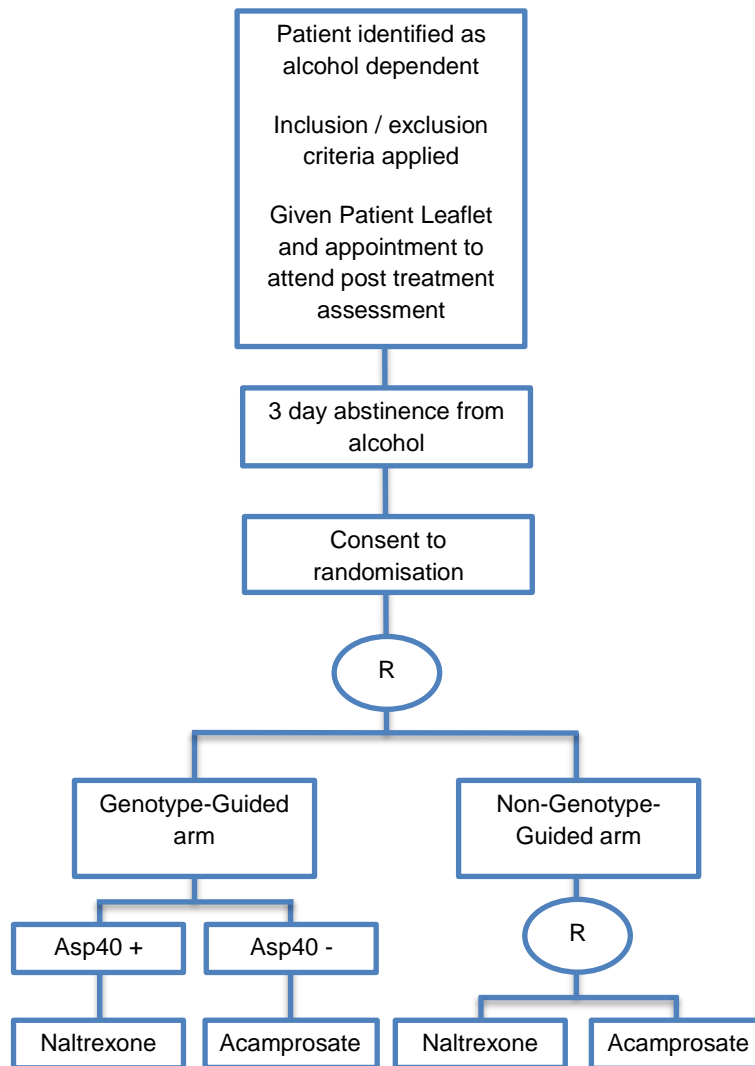
---

A number of studies, as described above, suggest a strong effect of the Asn40Asp SNP on naltrexone response in alcohol-dependent individuals. However, given the observed heterogeneity in effect sizes, and the fact that all studies conducted so far have explored the contribution of this SNP in a retrospective manner, and usually with small sample sizes, there is a need to ascertain the clinical efficacy conferred by genotype guided prescribing of naltrexone in alcohol dependent individuals in a prospective manner.

Recognizing this need, a randomized controlled trial was previously proposed to test the clinical utility of a stratified approach to naltrexone treatment. The proposed trial utilized a stratification design (which is called Biomarker Strategy Design [25]) proposed by the Institute of Medicine on their report on genomic biomarkers [26]. It was proposed that the design would allow the evaluation of whether genotype-guided prescribing of naltrexone based on the presence or absence of Asn40Asp SNP presented any clear benefit in terms of outcomes in alcohol-dependent patients.

The primary objective of the trial was to determine whether a genotype-guided treatment strategy would result in reduced rate of relapse to any drinking at 12 months in the treatment of alcohol dependence when compared to clinical care as recommended by NICE guidelines (non-genotype guided treatment strategy).

A two-armed, randomized controlled trial was proposed. In one arm (the non-genotype-guided arm) patients were to be randomized to either naltrexone or acamprosate (1:1 ratio). In the other arm (the genotype-guided arm), the treatment of patients was to be dependent on the carriage of the OPRM1 Asp40 allele. Patients with at least one copy of the Asp40 allele (patients homozygous or heterozygous for the Asp40 allele) would receive naltrexone whilst patients with no copy of the Asp40 allele (patients homozygous for the Asn40 allele) would receive acamprosate. The graphical illustration of the proposed design is given at Figure 6.1. Prescribing acamprosate in both randomisation arms in this way ensured that all patients received a pharmacotherapeutic intervention in accordance with NICE guidelines, to ensure that there was equipoise between both arms. In addition, despite the open label design of the trial, it ensured blinding to clinicians, nurses and patients in terms of the underlying genotype and reasons for choice of therapy.



**Figure 6.1.** Biomarker-strategy design with treatment randomization in the control arm adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and Acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.

Return to any drinking at 12 months was selected as the primary outcome which was also considered as the primary endpoint for most of the previous RCTs [7, 27]. It is a binary variable containing the information whether a patient returned to drinking after detoxification, or whether a patient remained fully abstinent.

The primary outcome would be measured utilizing a prospective drink diary. If at any stage during the 12 month follow-up period the patient recorded having a drink, they would be classified as having returned to drinking. ‘Event rate’ refers to



returning to drinking which is a negative outcome and ‘response rate’ refers to not returning to drinking which is a positive outcome.

### 6.2.2. Sample size calculation of the previously proposed design

---

**Sample size formula:** For the case of binary outcomes, Eng (2014) [28] provided formula (3.37) (see Subsection 3.2.4) for the required sample size for each arm in a Biomarker Strategy Design trial.

The expected response rates  $g_1$  in the genotype-guided arm and  $g_3$  in the non-genotype-guided arm can be found by calculating the formulae

$$g_1 = kr_{A+} + (1 - k)r_{B-} \quad (6.1)$$

and

$$g_3 = r_A/2 + r_B/2 \quad (6.2)$$

respectively;  $r_A$  and  $r_B$  denote the marginal effect of treatment A (naltrexone: experimental treatment) and treatment B (acamprosate: control treatment) respectively.  $r_{A+}$  and  $r_{B-}$  are the assumed response rates of biomarker-positive patients (i.e. patients with Asp40 present) receiving the experimental treatment and biomarker-negative patients (i.e. patients with Asp40 absent) receiving the control treatment respectively.  $k$  is the prevalence of biomarker-positive patients.

When designing the trial, the response rate in the non-genotype-guided arm ( $g_3$ ) is given, whereas, the response rate in the alternative arm ( $g_1$ ) is unknown. However, based on  $g_3$  and the decision made by a panel of clinicians regarding the minimally clinically meaningful effect size ( $\Delta_3$ ), we can calculate  $g_1$  by  $g_1 = \Delta_3 + g_3$ .

**Estimating expected response rate in non-genotype-guided arm,  $g_3$ :** Return to any drinking for acamprosate, without assuming dropouts returned to drinking, has an estimated event rate of 0.27 (95% CI: 0.23-0.32). This event rate was estimated from data on rate of return to any drinking and rate of dropouts from two meta-

analyses (analysis 1.1. and 1.16 respectively) of the Cochrane acamprosate review [5]. Results are illustrated in Figure 6.2.

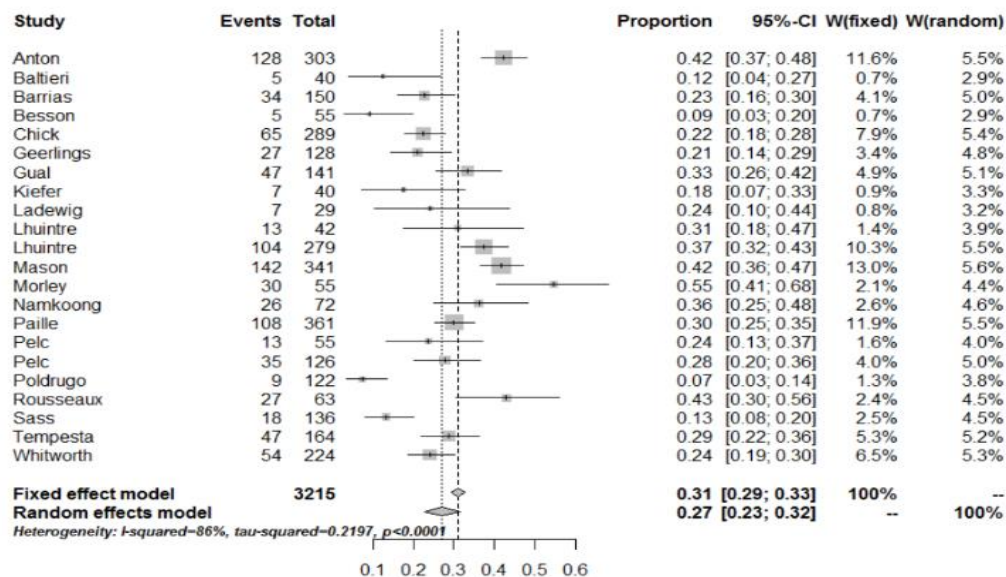


Figure 6.2. Rate of return to any drinking for acamprosate

The COMBINE study [7] has reported a difference of 3% (95% CI: - 4%, 9%) between acamprosate and naltrexone event rates of those having returned to drinking at 1 year. Taking the upper limit of the 95% CI for the estimated acamprosate event rate (0.32) i.e. assuming the worst case scenario, and the 3% difference estimated in the COMBINE study, return to any drinking for naltrexone has an estimated event rate of 0.29. The overall event rate in the non-genotype-guided arm ( $g_3$ ) can therefore be expected to be 0.31 as calculated by  $0.5 \times 0.32$  (acamprosate) +  $0.5 \times 0.29$  (naltrexone) based on (6.2). Consequently, the response rate in the baseline arm (i.e. non-genotype-guided arm) can be estimated as  $g_3 = 1 - 0.31 = 0.69$ .

**Estimating expected response rate in genotype-guided arm,  $g_1$ :** In this arm, as patients who are positive for the Asp40 genotype receive only naltrexone and patients who are negative for the Asp40 genotype receive only acamprosate, the anticipated event rate is a weighted event rate of Asp40 genotype positive patients receiving naltrexone and Asp40 genotype negative patients receiving acamprosate. Therefore, based on data on rate of return to any drinking in patients on naltrexone carrying at least one copy of the Asp40 allele (12.9%) from the COMBINE genetics

sub-study [11], and given a 20% prevalence of the Asp40 allele, combining this with the upper limit of the event rate estimate from the Cochrane acamprosate review, the best estimate for the event rate in the genotype arm can be calculated using (6.1) by  $0.2$  (i.e. biomarker positive prevalence)  $\times 0.129$  (i.e. biomarker positive naltrexone event rate)  $+ 0.8$  (i.e. biomarker negative prevalence)  $\times 0.32$  (i.e. acamprosate event rate)  $= 0.282$ .

Consequently, the response rate in the genotype-guided arm would be  $1 - 0.282 = 0.72$ . The  $0.72$  response rate in the genotype arm leads to a difference in response rates between non-genotype and genotype-guided arms ( $\Delta_3$ ) of  $0.72 - 0.69 = 0.03$  (3% difference), and this estimate provides us with an estimate of what a realistic effect size might be. Even in the extreme situation where the 20% with the Asp40 allele have no relapse, the event rate in the genotype arm can be calculated using (6.1) by  $0.2 \times 0 + 0.8 \times 0.32 = 0.256$ . Hence, the response rate in the genotype-guided arm would be  $1 - 0.256 = 0.74$ . The  $0.74$  response rate in the genotype arm leads to a difference in response rates between non-genotype and genotype-guided arms ( $\Delta_3$ ) of  $0.74 - 0.69 = 0.05$  (5% difference). For this reason, realistically the difference in outcomes between the two arms based on data from the literature is known to be around 5% or less. Consequently, we will consider in our case study three scenarios of target effect sizes ( $\Delta_3$ ), 3%, 5% and 10% for the sample size calculation. After having set the values of  $\Delta_3$ , the response rate in the genotype-guided arm ( $g_1$ ) can be found by  $g_1 = \Delta_3 + g_3$ .

In reality, the assumed effect sizes should be discussed with a panel of expert clinicians who will decide which effect size can be considered as a minimally clinically significant difference.

Based on the aforementioned effect sizes and the response rate in the non-genotype-guided arm we calculate the corresponding sample size and the results are presented in Table 6.1.

**Table 6.1.** Sample size of the Biomarker-strategy design with treatment randomisation in the control arm in each effect size scenario.

	Scenario 1	Scenario 2	Scenario 3
Non-genotype guided arm response rate ( $g_3$ )	0.69	0.69	0.69
Difference in response rates between strategy arms ( $\Delta_3$ )	0.03 (3%)	0.05 (5%)	0.1 (10%)
Genotype-guided arm response rate ( $g_1 = \Delta_3 + g_3$ )	0.72	0.74	0.79
Two-sided significance level ( $\alpha$ )	0.05	0.05	0.05
Power ( $1 - \beta$ )	0.8	0.8	0.8
Sample size per arm calculated by (3.37)	3624	1276	298
Sample size per arm allowing for a 30% dropout rate calculated by [(3.37)/(1 – dropout rate)]	5178	1824	426
Total sample size	10356	3648	852

To demonstrate a 3% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and at a two-sided 5% significance level, 3624 patients will be required per arm. Allowing for a 30% dropout rate, 5178 patients will be required per arm, therefore a total of 10356 patients will need to be recruited. To demonstrate a 5% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and a two-sided significance level, 1276 patients will be required per arm. Allowing for a 30% dropout rate, 1824 patients will be required per arm, therefore a total of 3648 patients will be recruited. To demonstrate a 10% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and at a two-sided 5% significance level, 298 patients will be required per arm. Allowing for a 30% dropout rate, 426 patients will be required per arm, therefore a total of 852

patients will be recruited. However, according to data identified within the literature review, the 10% difference in response rates between the strategy arms is unlikely.

### 6.3. Reasons for inefficiency of the previously proposed design

---

This Biomarker-strategy design with treatment randomization in the control arm is inefficient in our case as a very large sample size is required. This design is generally a poor substitute for clinical trials which aim to compare the experimental treatment to control treatment, since we have biomarker-positive and biomarker-negative patients in both the biomarker-guided arm and non-biomarker-guided arm being assigned to the same treatment, therefore diluting the treatment effect [29, 30]. Consequently, as a large overlap of patients receiving the same treatment might occur, the comparison of the two biomarker-strategy arms results in an odds ratio which is forced towards unity, i.e. no treatment effect [29]. For this reason, a large sample size is needed especially in cases where the prevalence of biomarker is low and the assumed overall difference between the two biomarker-strategy arms is small [29].

Generally, the Biomarker-Strategy Design with treatment randomization in the control arm needs a larger sample size as compared to so-called ‘marker-stratified’ designs and for this reason it is considered a less efficient design [29]. However, a strength of this design is that it allows clarification of whether the results which indicate efficacy of the biomarker-guided approach to treatment are caused due to a true effect of the biomarker or due to a treatment effect irrespective of biomarker status.

As this design has been proven inefficient for our study we reconsider in the next section what might be the most appropriate design among the various non-adaptive biomarker-guided clinical trial designs outlined in Chapter 3.

## 6.4. Reconsideration of the most appropriate design

---

By taking into consideration the information regarding the utility of each design given in Chapter 3 [29], we extracted the following questions which we believed were key when deciding which non-adaptive design to use for the STRONG trial:

### *1. Is our goal to test a treatment effect in a biomarker-positive subgroup only?*

This aim would be appropriate where there was prior evidence indicating that effectiveness was limited to those within that particular subgroup, but the candidate biomarker still required prospective validation [29], and is the aim being addressed in the ‘Enrichment’ trial design. The design is recommended when the analytical validity of the biomarker has been well established and is particularly appropriate where it is unethical to randomize the biomarker-negative population into different treatment arms, for example where there is prior evidence that the experimental treatment is not beneficial for biomarker-negative individuals, or is likely to cause them harm. However, when it remains unclear whether or not biomarker-negative individuals will benefit from the novel treatment, the enrichment design is not appropriate and alternative designs, which also assess effectiveness in the biomarker-negative individuals, should be considered [29].

**STRONG trial:** Our aim is to test treatment effectiveness both in biomarker-positive and biomarker-negative patients. Current NICE guidelines (CG115) [6] recommends therapy with either acamprosate (control treatment in STRONG) or naltrexone (experimental treatment in STRONG) to alcohol dependent patients (regardless of genotype), hence, both genotype groups should be investigated in our study and the Enrichment design is therefore not appropriate.

### *2. Is our biomarker extremely rare?*

If the biomarker is rare (<20%), then the Enrichment design should again be considered. An all-comers strategy is not recommended in this case as the treatment effect in the overall population would be diluted [29].

**STRONG trial:** In our study the prevalence of Asn40Asp SNP is 20%. This proportion is low; however, it is not extremely rare.

*3. Is our goal to test treatment effectiveness both in biomarker-positive and biomarker-negative patients as well as to assess the interaction between treatment and biomarker in all patients?*

A design which allows all these values to be estimated is appropriate when there is enough evidence that the experimental treatment is more effective in the biomarker-positive subgroup than in the biomarker-negative subgroup, but when there is insufficient evidence that the experimental treatment is of no benefit in biomarker-negative individuals [29]. Such designs are referred to as 'Marker-stratified' designs, of which there are several variations including 'Subgroup-specific' designs, 'Biomarker-positive and overall strategies' design, and 'Marker sequential test' design ('MAST') (see Chapter 3, Subsection 3.2.3.1.) [29]. Marker-stratified designs can be conducted using two different analysis plans; the so-called 'marker-by-treatment interaction with separate tests' and 'marker-by-treatment interaction with interaction test' (see Chapter 3, Subsection 3.2.3.1.) [29].

**STRONG trial:** The aim of the STRONG trial is to test effectiveness within both biomarker-defined subgroups as well as to investigate the interaction term which will indicate whether the biomarker is predictive or not. Hence, the Marker-stratified design could be a potential design for our trial.

*4. Is our goal to test a treatment effect in a biomarker-positive subgroup only but there is compelling prior evidence which shows efficacy of the control treatment for the biomarker-negative subgroup?*

Where this is the case, Hybrid designs are recommended. Hybrid designs are particularly recommended when there is compelling prior evidence showing a detrimental effect of the experimental treatment for the biomarker-negative subgroup, or some indication of its possible excessive toxicity, as well as evidence showing that the control treatment works well in that subgroup [29].

**STRONG trial:** There is no evidence of our experimental treatment, naltrexone, having a detrimental effect in Asp40- patients, so this design is inappropriate for our trial.

#### *5. Is our goal to test a clinical strategy in all patients?*

Designs aimed at testing the clinical strategy of offering a biomarker-guided approach to treatment take into consideration the prevalence of the biomarker as well as its interaction with treatment to characterize the potential impact of the biomarker strategy in clinical practice care. Such designs are called ‘Biomarker-strategy’ designs and test the joint deployment of the biomarker and active treatment as a strategy. They include the following designs: (i) biomarker-strategy with biomarker assessment in the control arm, (ii) biomarker-strategy without biomarker assessment in the control arm, (iii) biomarker-strategy with treatment randomization in the control arm and (iv) reverse marker-based strategy designs [29].

**STRONG trial:** Since the aim of our trial is to test the effectiveness of a genotype-guided approach to treatment of alcohol-dependent patients, a Biomarker-strategy design may be appropriate.

So far, we have therefore determined that either a Marker-stratified design or a Biomarker-strategy design may be appropriate for the STRONG trial. The following questions help us further to choose among the different biomarker-strategy designs.

#### *6. Is our goal to test the hypothesis that the treatment effect based on the biomarker-guided strategy approach is superior to that of the standard of care where all patients receive only the control treatment?*



If this is the case, either the Biomarker-strategy design with biomarker assessment in the control arm or the Biomarker-strategy design without biomarker assessment in the control arm could be used. The latter would be preferred in situations where it is either not feasible or unethical to test the biomarker in the entire population (e.g., specimens not submitted, assay failure, tumour inaccessible) [29].

**STRONG trial:** Since NICE guidance (CG115) [6] recommends therapy with either acamprosate (experimental treatment) or naltrexone (control treatment), neither treatment can be considered the ‘standard of care’ approach and therefore these two designs cannot be considered for the STRONG trial.

*7. Is our goal to test a clinical strategy in all patients as well as to test whether the biomarker is not only prognostic but also predictive (interaction between treatment and biomarker)?*

The Biomarker-strategy design with treatment randomization in the control arm is appropriate when this is the goal. Patients are first randomly assigned to either the genotype-guided strategy arm or to the non-genotype-guided strategy arm. Next, patients who are allocated to the non-genotype-guided strategy are again randomized either to the experimental treatment arm or to the standard treatment arm irrespective of their biomarker status. Patients who are allocated to the genotype-guided strategy and who are biomarker-positive receive the experimental treatment and patients who are biomarker-negative receive the control treatment [29].

**STRONG trial:** In our trial, we are aiming to test the effectiveness of a genotype-guided treatment strategy, and are wishing to test whether the SNP of interest is predictive and so this is a potential design for our study. However, this is the design which we previously considered and found that it was not plausible due to the relatively low prevalence of the genetic marker.

*8. Is our aim to test the clinical strategy of a biomarker-guided approach in all patients and the predictive ability of the biomarker, whilst ensuring that the*

*probability of a patient being assigned the same treatment regardless of which arm they are randomized to is zero?*

The Reverse Marker-Based strategy design is appropriate where this is the case. Patients are randomized either to the genotype-guided strategy arm or to a reverse genotype-guided strategy arm. Patients who are allocated to the genotype-guided arm receive the experimental treatment if they are biomarker-positive whereas biomarker-negative patients receive the control treatment. By contrast, patients who are randomly assigned to the reverse genotype-guided arm receive control treatment if they are biomarker-positive, whereas biomarker-negative patients receive the experimental treatment.

**STRONG trial:** Since we are interested in testing the effectiveness of our genotype-guided approach to treatment and whether our SNP is predictive, and at the same time have prior evidence that both naltrexone and acamprosate are effective in treating alcohol dependence (hence current NICE guidance), it is ethical to adopt this design in our trial.

## 6.5. Sample size calculations for the STRONG trial assuming different study designs

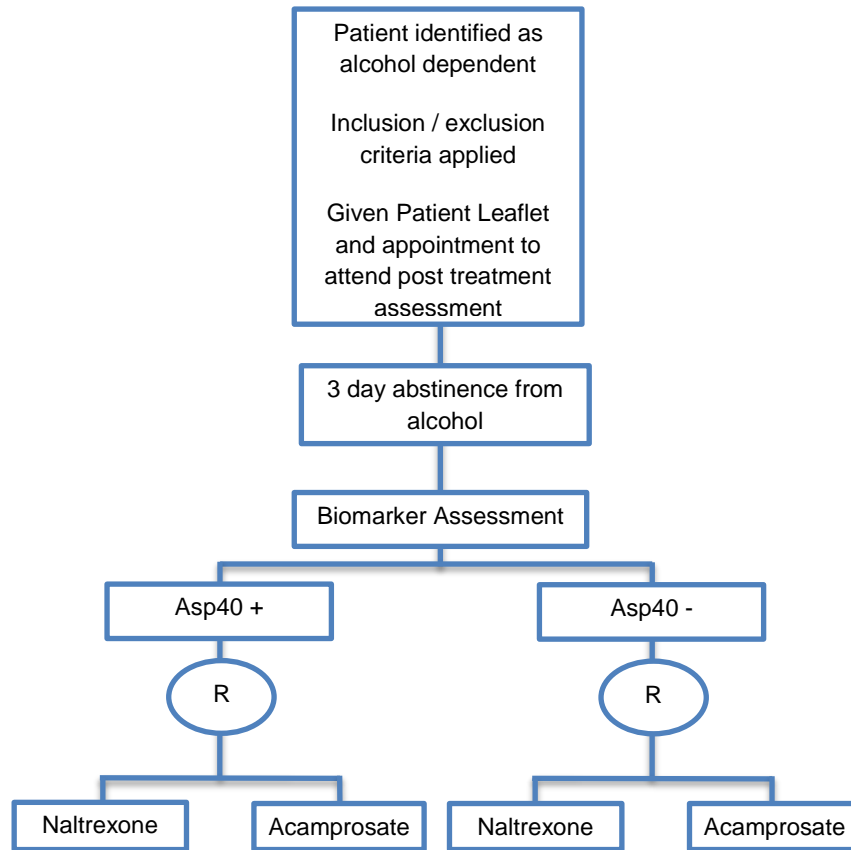
---

Based on our conclusions above, in the following sections we investigate the use of two biomarker-guided trial designs in the STRONG trial i.e. the Marker Stratified design (two variations of that design are also included) and the Reverse Marker-Based strategy design. The primary outcome, upon which our sample size calculation is based, is return to any drinking at 12 months.

### 6.5.1. Using the Marker Stratified design in the STRONG trial

---

The graphical representation of this design is given in Figure 6.3.



**Figure 6.3.** Marker Stratified design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.

**Sample size formula:** Eng (2014) [28] provided formula (3.20) for the required total sample size for a trial with this design, assuming a binary outcome (see Subsection 3.2.3.1.).

$r_{B+}$ ,  $r_{A-}$  are the assumed response rates of biomarker-positive patients (i.e. patients who carry the Asp40 ) receiving the control treatment and biomarker-negative patients i.e. patients who do not carry the Asp40 allele) receiving the experimental treatment. The sample size for the Marker Stratified design does not depend on the biomarker prevalence, therefore the trial arm can be closed once the required number of patients has accrued, without the need to re-estimate the sample size. Of course, the lower the prevalence the longer the accrual period for the biomarker positive arm. For an actual clinical study which uses this design, the

sample size calculation is based on the response rate in the control arm of each biomarker-defined subgroup (i.e.  $r_{B+}$ ,  $r_{B-}$ ) and on the minimally clinically meaningful effect size of each biomarker-defined subgroup (i.e.  $r_{A+} - r_{B+}$  and  $r_{A-} - r_{B-}$ ) which is decided by a panel of clinicians. Based on these parameters, the response rate in the experimental arm of each biomarker-defined subgroup can be found and thus the sample size of each subgroup can be calculated.

As mentioned in the previous section regarding the biomarker-strategy design, the event rate for return to any drinking for acamprosate can be estimated as from the Cochrane acamprosate review [5] to be 0.32. Consequently, the estimated response rate for those on acamprosate is estimated as  $1 - 0.32 = 0.68$ . If we consider the impact of genetics, based on data on rates of return to any drinking in patients on naltrexone carrying at least one copy of the Asp40 allele (12.9%) and those not carrying the Asp40 allele (45.2%) from the COMBINE genetics sub-study [11], the estimated naltrexone response rate in the biomarker-positive subgroup is  $1 - 0.129 = 0.871$  and in the biomarker-negative subgroup is  $1 - 0.452 = 0.548$ . We have identified no studies investigating association between acamprosate response and the Asp40Asn SNP. Hence, we will assume that the acamprosate response rate is independent of genotype at the SNP and is therefore 0.68. Based on this data, the difference in response rates between the naltrexone and acamprosate treatment arms in the biomarker-positive subgroup can be estimated as  $0.871 - 0.68 = 0.191$  and the difference in the biomarker-negative subgroup can be estimated as  $0.548 - 0.68 = -0.132$  which indicates that the acamprosate treatment performs better compared to the naltrexone treatment.

These estimated differences provide us with an idea of what the true differences may be in each of the biomarker subgroups. In reality, the differences would be discussed with a panel of clinicians in order to decide which effect size can be considered as a minimally clinically significant difference for each subgroup. However, in our sample size calculation we assume three scenarios of differences with a range from small difference to high difference, and the results are presented in Table 6.2; scenarios 4 to 9 give some combinations of the first three scenarios. The

first scenario corresponds to differences of 5% and -20% in biomarker-positive subgroup and biomarker-negative subgroup respectively. The second scenario corresponds to differences of 10% and -15% in biomarker-positive subgroup and biomarker-negative subgroup respectively. The third scenario corresponds to differences of 19.1% and -13.2% in biomarker-positive and biomarker-negative subpopulation respectively, in line with our estimates based on published data. Sample sizes, based on these assumptions, are provided in Table 6.2.

**Table 6.2.** Sample size of the Marker Stratified design based on the target effect size and acamprosate response rate in each biomarker-defined subgroup.

	Scenarios								
	1	2	3	4	5	6	7	8	9
Acamprosate response rate in biomarker-positive subgroup ( $r_{B+}$ )	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
Acamprosate response rate in biomarker-negative subgroup ( $r_{B-}$ )	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
Difference in response rates between naltrexone and acamprosate treatment arm in biomarker-positive subgroup (effect size in positive subgroup)	0.05	0.10	0.191	0.05	0.05	0.10	0.10	0.191	0.191
Difference in response rates between naltrexone and acamprosate treatment arm in biomarker-negative subgroup (effect size in negative subgroup)	-0.20	-0.15	-0.132	-0.15	-0.132	-0.20	-0.132	-0.20	-0.15
Naltrexone response rate in biomarker-positive subgroup  ( $r_{A+}$ = effect size in positive subgroup + $r_{B+}$ )	0.73	0.78	0.871	0.73	0.73	0.78	0.78	0.871	0.871
Naltrexone response rate in biomarker-negative subgroup  ( $r_{A-}$ = effect size in negative subgroup + $r_{B-}$ )	0.48	0.53	0.548	0.53	0.548	0.48	0.548	0.48	0.53
Two-sided significance level ( $\alpha$ )	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

<b>Power (<math>1 - \beta</math>)</b>	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
<b>Total sample size (N) calculated by (3.20)</b>	2787	937	561	2930	3023	794	1030	325	468
<b>Total sample size allowing for a 30% dropout rate calculated by [(3.20)/(1 – dropout rate)]</b>	3982	1339	801	4186	4319	1134	1472	465	669

In the first scenario of differences, with 80% power and at a 5% significance level, 2787 patients will be required. Allowing for a 30% dropout rate, 3982 patients will need to be recruited if this study design is assumed. In the second scenario of differences, with 80% power and at a 5% significance level, 937 patients will be required. Allowing for a 30% dropout rate, 1339 patients will need to be recruited if this study design is assumed. In the third scenario of differences, with 80% power and at a 5% significance level, 561 patients will be required. Allowing for a 30% dropout rate, 801 patients will need to be recruited if this study design is assumed. In the fourth scenario of differences, with 80% power and at a 5% significance level, 2930 patients will be required. Allowing for a 30% dropout rate, 4186 patients will need to be recruited if this study design is assumed. In scenario 5 of differences, with 80% power and at a 5% significance level, 3023 patients will be required. Allowing for a 30% dropout rate, 4319 patients will need to be recruited if this study design is assumed. In scenario 6 of differences, with 80% power and at a 5% significance level, 794 patients will be required. Allowing for a 30% dropout rate, 1134 patients will need to be recruited if this study design is assumed. In scenario 7 of differences, with 80% power and at a 5% significance level, 1030 patients will be required. Allowing for a 30% dropout rate, 1472 patients will need to be recruited if this study design is assumed. In scenario 8 of differences, with 80% power and at a 5% significance level, 325 patients will be required. Allowing for a 30% dropout rate, 465 patients

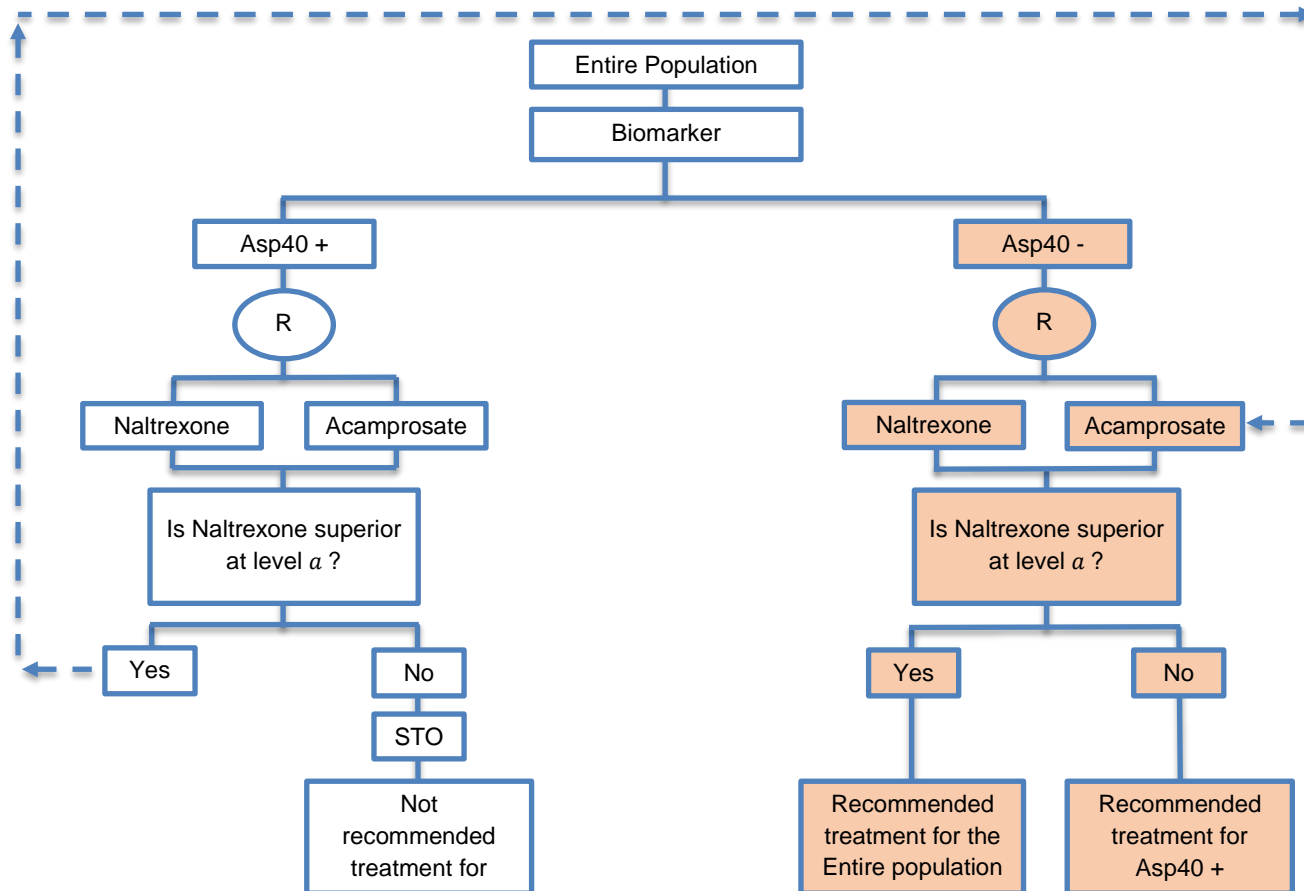
will need to be recruited if this study design is assumed. In scenario 9 of differences, with 80% power and at a 5% significance level, 468 patients will be required. Allowing for a 30% dropout rate, 669 patients will need to be recruited if this study design is assumed.



#### *6.5.1.1. Using a variation of the Marker Stratified design – the Sequential Subgroup-Specific design*

---

In this section we investigate a version of Marker Stratified design, the so-called Sequential Subgroup-Specific design (see Chapter 3, Subsection 3.2.3.1), which consists of the analysis of biomarker-negative patients contingent on statistical significance in biomarker-positive patients. A graphical illustration of this design is given in Figure 6.4.



**Figure 6.4.** Sequential Subgroup Specific design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.

**Sample size formula:** Simon (2008) [31] provided formula (3.28) for the total sample size for this trial design, again assuming a binary outcome (see Subsection 3.2.3.1.). This formula is composed of the sum of the formula (3.27) which corresponds to the biomarker-positive patients and (3.29) which corresponds to the biomarker-negative patients. These formulae depend on  $N_{enrichment}$  which denotes the required total sample size for a trial assuming an enrichment design [29]. By using the part of formula (3.20) (see Subsection 3.2.3.1.) proposed in Eng (2014) [28] which refers to the biomarker-positive subgroup, formula (3.28) can be written as:

$$N = 2(z_\alpha + z_{1-\beta})^2 \left\{ \frac{r_{A+}(1 - r_{A+}) + r_{B+}(1 - r_{B+})}{(r_{A+} - r_{B+})^2} \right\} \quad (6.3)$$

$$+ 2 \frac{(1 - k)}{k} (z_\alpha + z_{1-\beta})^2 \left\{ \frac{r_{A+}(1 - r_{A+}) + r_{B+}(1 - r_{B+})}{(r_{A+} - r_{B+})^2} \right\}$$

For the purpose of our sample size calculation, we make the same assumptions regarding the response rate in the biomarker-positive subgroup and the effect size as we did when considering the Marker Stratified design. The sample size calculation is given in Table 6.3. The same scenarios of difference in response rates between naltrexone and acamprosate treatment arm in biomarker-positive subgroup used in the Marker Stratified design are considered.

**Table 6.3.** Sample size of the Sequential Subgroup-Specific design based on the target effect size and acamprosate response rate in biomarker-positive subgroup.

	Scenario 1	Scenario 2	Scenario 3
<b>Acamprosate response rate in biomarker-positive subgroup (<math>r_{B+}</math>)</b>	0.68	0.68	0.68
<b>Difference in response rates between naltrexone and acamprosate treatment arm in biomarker-positive subgroup (effect size in positive subgroup)</b>	0.05	0.10	0.191

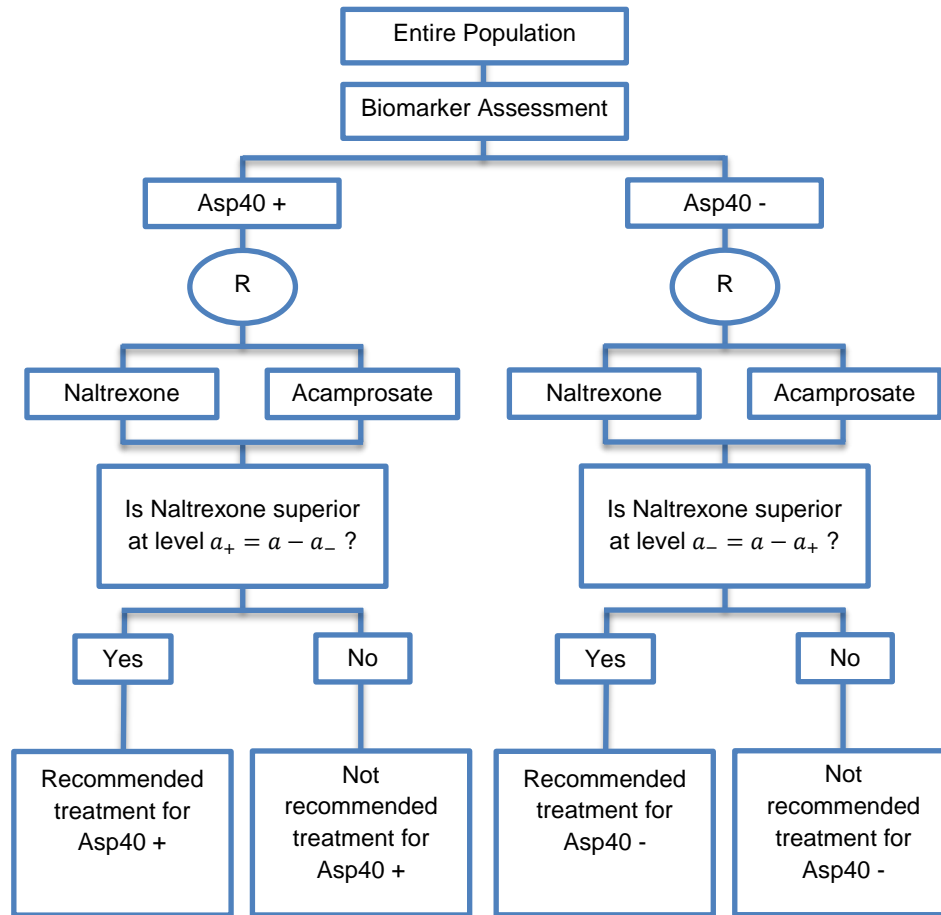
<b>Naltrexone response rate in biomarker-positive subgroup (<math>r_{A+}</math> = effect size in positive subgroup + <math>r_{B+}</math>)</b>	0.73	0.78	0.871
<b>Two-sided significance level (<math>\alpha</math>)</b>	0.05	0.05	0.05
<b>Power (<math>1 - \beta</math>)</b>	0.8	0.8	0.8
<b>Total sample size (N) calculated by (6.3)</b>	13019	3055	710
<b>Total sample size allowing for a 30% dropout rate calculated by <math>[(6.3)/(1 - \text{dropout rate})]</math></b>	18598	4364	1014

In scenario 1, with 80% power and at a 5% significance level, 13019 patients will be required. Allowing for a 30% dropout rate, 18598 patients will need to be recruited if this study design is adopted. In scenario 2, with 80% power and at a 5% significance level, 3055 patients will be required. Allowing for a 30% dropout rate, 4363 patients will need to be recruited if this study design is adopted. In scenario 3, with 80% power and at a 5% significance level, 710 patients will be required. Allowing for a 30% dropout rate, 1014 patients will need to be recruited if this study design is adopted.

#### 6.5.1.2. Using a variation of the Marker Stratified design: the Parallel Subgroup-Specific design

---

In this section another version of the Marker Stratified design is explored, the so-called Parallel Subgroup-Specific design which tests both biomarker-defined subgroups, however it controls for the type I error (see Chapter 3, Subsection 3.2.3.1.). The graphical illustration of this design is given in Figure 6.5.



**Figure 6.5.** Parallel Subgroup-Specific design adopted in STRONG trial. “R” refers to randomisation of patients. Naltrexone corresponds to the experimental treatment (A) and acamprosate corresponds to the control treatment (B). Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.

**Sample size formula:** When more than one hypothesis for treatment efficacy is being tested, it is important to control the familywise error rate (FWER) to ensure that the probability of committing at least one Type I error does not exceed the nominal significance level. To achieve this within the current study design, a conservative Bonferroni correction method is often used where the nominal significance level  $\alpha$  is allocated between the test for the biomarker-negative subgroup and the test for the biomarker-positive subgroup either equally (i.e.  $\alpha/2$ ) or unequally, meaning that the total significance level is  $\alpha$ . The following sample size formula assuming  $\alpha_+$  and  $\alpha_-$  significance levels for biomarker-positive and biomarker-negative subgroups respectively was proposed in Eng (2014) [28] for this design,

$$\begin{aligned}
N = & 2(z_{a_+} + z_{1-\beta})^2 \left\{ \frac{r_{A+}(1 - r_{A+}) + r_{B+}(1 - r_{B+})}{(r_{A+} - r_{B+})^2} \right\} \\
& + 2(z_{a_-} + z_{1-\beta})^2 \left\{ \frac{r_{A-}(1 - r_{A-}) + r_{B-}(1 - r_{B-})}{(r_{A-} - r_{B-})^2} \right\}.
\end{aligned} \tag{6.4}$$

For the purpose of our sample size calculation, we again make the same assumptions regarding the response rate in the biomarker-positive subgroup and the effect size as we did when considering the Marker Stratified design. The sample size calculation is given in Table 6.4; scenarios 4 to 9 give some combinations of the first three scenarios. The same scenarios of difference in response rates between naltrexone and acamprosate treatment arm in biomarker-positive subgroup and biomarker-negative subgroup used in the Marker Stratified design are considered.

**Table 6.4.** Sample size of the Parallel Subgroup-Specific design based on the target effect size and acamprosate response rate in each biomarker-defined subgroup.

	Scenarios								
	1	2	3	4	5	6	7	8	9
Acamprosate response rate in biomarker-positive subgroup ( $r_{B+}$ )	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
Acamprosate response rate in biomarker-negative subgroup ( $r_{B-}$ )	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
Difference in response rates between naltrexone and acamprosate treatment arm in biomarker-positive subgroup (effect size in positive subgroup)	0.05	0.10	0.191	0.05	0.05	0.10	0.10	0.191	0.191
Difference in response rates between naltrexone and acamprosate treatment arm in biomarker-negative subgroup (effect size in negative subgroup)	-0.20	-0.15	-0.132	-0.15	-0.132	-0.20	-0.132	-0.20	-0.15
Naltrexone response rate in biomarker-positive subgroup ( $r_{A+}$ = effect size in positive subgroup + $r_{B+}$ )	0.73	0.78	0.871	0.73	0.73	0.78	0.78	0.871	0.871

<b>Naltrexone response rate in biomarker-negative subgroup</b> ( $r_{A-}$ = effect size in negative subgroup + $r_{B-}$ )	0.48	0.53	0.548	0.53	0.548	0.48	0.548	0.48	0.53
<b>Two-sided significance level (<math>\alpha_+</math>, <math>\alpha_-</math>)</b>	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025	0.025
<b>Power (<math>1 - \beta</math>)</b>	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
<b>Total sample size (N) calculated by (6.4)</b>	3375	1134	680	3547	3661	962	1248	394	566
<b>Total sample size allowing for a 30% dropout rate</b> calculated by [(6.4)/(1 – dropout rate)]	4821	1620	971	5067	5230	1374	1782	563	809



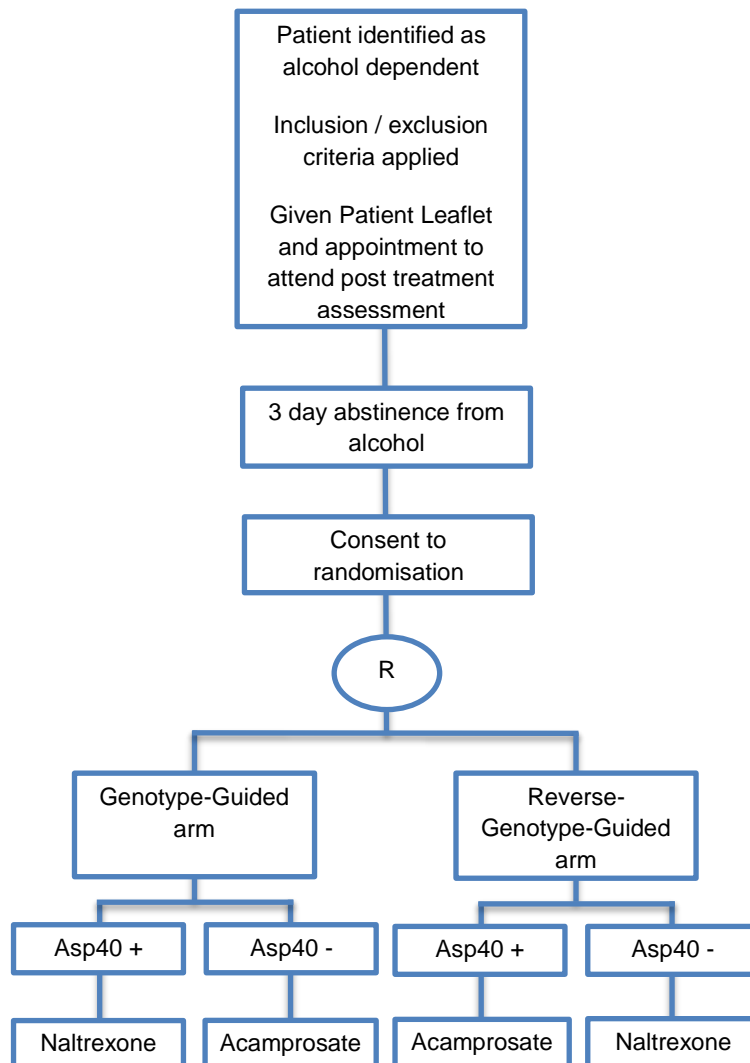
In scenario 1, with 80% power and at a 5% significance level, 3375 patients will be required. Allowing for a 30% dropout rate, 4821 patients will need to be recruited if this study design is chosen. In scenario 2, with 80% power and at a 5% significance level, 1134 patients will be required. Allowing for a 30% dropout rate, 1620 patients will need to be recruited if this study design is chosen. In scenario 3, with 80% power and at a 5% significance level, 680 patients will be required. Allowing for a 30% dropout rate, 971 patients will need to be recruited if this study design is chosen. In scenario 4, with 80% power and at a 5% significance level, 3547 patients will be required. Allowing for a 30% dropout rate, 5067 patients will need to be recruited if this study design is chosen. In scenario 5, with 80% power and at a 5% significance level, 3661 patients will be required. Allowing for a 30% dropout rate, 5230 patients will need to be recruited if this study design is chosen. In scenario 6, with 80% power and at a 5% significance level, 962 patients will be required. Allowing for a 30% dropout rate, 1374 patients will need to be recruited if this study design is chosen. In scenario 7, with 80% power and at a 5% significance level, 1248 patients will be required. Allowing for a 30% dropout rate, 1782 patients will need to be recruited if this study design is chosen. In scenario 8, with 80% power and at a 5% significance level, 394 patients will be required. Allowing for a 30% dropout rate, 563 patients will need to be recruited if this study design is chosen. In scenario 9, with 80% power and at a 5% significance level, 566 patients will be required. Allowing for a 30% dropout rate, 809 patients will need to be recruited if this study design is chosen.

#### 6.5.2. Using the Reverse Marker-Based strategy design in the STRONG trial

---

A graphical illustration of this approach is given in Figure 6.6. Patients randomized to the reverse genotype-guided arm will receive only acamprosate (control treatment) if they carry the Asp40 allele and those who do not carry the Asp40 allele will receive only naltrexone (experimental treatment). For the genotype guided arm, patients who carry the Asp40 allele will receive naltrexone (experimental

treatment) while patients who do not carry the Asp40 allele will receive acamprosate (control treatment).



**Figure 6.6.** Reverse Marker-Based strategy design adopted in STRONG trial. “R” refers to randomization of patients. Naltrexone corresponds to the experimental treatment and acamprosate corresponds to the control treatment. Asp40 + (Asp40 present) corresponds to biomarker-positive patients and Asp40 - (Asp40 absent) corresponds to biomarker-negative patients.

**Sample size formula:** For the case of binary outcomes, Eng (2014) [28] provided formula (3.38) for the required sample size for each arm in a trial with this design (see Subsection 3.2.4.4.).

The expected response rates  $g_1$ ,  $g_4$  in the genotype-guided arm and in the reverse-genotype-guided arm respectively can be found by calculating the formulae

$$kr_{A+} + (1 - k)r_{B-} \quad (6.5)$$

and

$$kr_{B+} + (1 - k)r_{A-} \quad (6.6)$$

For an actual clinical study which uses this design, the sample size calculation should be based on the response rate in the control arm, i.e. the response rate in the reverse-genotype-guided arm ( $g_4$ ) and on the decision made by a panel of clinicians regarding the minimally clinically meaningful effect size, i.e.  $\Delta_4$ . Hence, the response rate in the alternative arm can be found by  $g_1 = \Delta_4 + g_4$ .

Based on the data obtained from the literature search that has been undertaken we can estimate the response rate in the genotype-guided arm and in the reverse-genotype-guided arm and thus what may be a realistic effect size that we can expect to observe. As discussed in the Biomarker-strategy design, the response rate on acamprosate can be assumed as 0.32, based on the Cochrane acamprosate review [5]. If we consider the impact of genetics, we can estimate rates of return to any drinking in patients on naltrexone carrying at least one copy of the Asp40 allele (12.9%) and those not carrying the Asp40 allele (45.2%) from the COMBINE genetics sub-study [11]. We have not identified any studies investigating association between the Asp40Asn SNP and response to acamprosate hence we assume that the acamprosate response rate in both the biomarker-positive and biomarker-negative subgroups is 0.32. Assuming a 20% prevalence of the Asp40 allele and using formula (6.6) above, the event rate in the reverse-genotype arm is calculated to be 0.43 [ $0.2 \times 0.32$  (i.e. acamprosate) +  $0.8 \times 0.452$  (i.e. naltrexone in negative subgroup)]. Consequently, the response rate in the reverse-genotype-guided arm is estimated as  $g_4 = 1 - 0.43 = 0.57$ . Further, using formula (6.5) above, the overall event rate in the genotype-guided arm

is expected to be  $0.28 [0.2 \times 0.129 \text{ (i.e. naltrexone in positive subgroup)} + 0.8 \times 0.32 \text{ (i.e. acamprosate)}]$ . Hence, the response rate in the genotype-guided arm is estimated as  $g_1 = 1 - 0.28 = 0.72$ .

Based on published data, the difference in response rates between the reverse-genotype-guided arm and the genotype-guided arm can therefore be estimated as  $0.72 - 0.57 = 0.15$  (15%). Even if the 20% with the Asp40 genotype have no relapse the difference in response rates would be  $[0.2 \times (1 - 0) + 0.8 \times (1 - 0.32)] - 0.57 = 0.17$  (17%). For this reason, realistically the difference in outcomes between the two arms of the trial based on data from the literature is known to be around 17% or less. In our case study we consider three scenarios of target effect sizes ( $\Delta_4$ ) with a range from small effect size to higher effect size, 15%, 17%, 20% for the sample size calculation. After having set the aforementioned values of  $\Delta_4$ , the response rate in the genotype-guided arm ( $g_1$ ) can be found by  $g_1 = \Delta_4 + g_4$ . In reality, these scenarios of effect sizes should be discussed with a panel of expert clinicians who will decide which effect size can be considered as a minimally clinically significant difference. Results of estimating the sample size under these scenarios are presented in Table 6.5.

**Table 6.5.** Sample size of the Reverse Marker-Based strategy design with treatment randomization in the control arm in each effect size scenario.

	Scenario 1	Scenario 2	Scenario 3
<b>Reverse-genotype guided arm response rate (<math>g_4</math>)</b>	0.57	0.57	0.57
<b>Difference in response rates between strategy arms (<math>\Delta_4</math>)</b>	0.15 (15%)	0.17 (17%)	0.20 (20%)
<b>Genotype-guided arm response rate (<math>g_1 = \Delta_4 + g_4</math>)</b>	0.72	0.74	0.77

<b>Two-sided significance level (<math>\alpha</math>)</b>	0.05	0.05	0.05
<b>Power (<math>1 - \beta</math>)</b>	0.8	0.8	0.8
<b>Sample size per arm calculated by (3.38)</b>	156	120	83
<b>Sample size per arm allowing for a 30% dropout rate calculated by <math>[(3.38)/(1 - \text{dropout rate})]</math></b>	224	172	119
<b>Total sample size</b>	448	344	238

To demonstrate a 15% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and at a two-sided 5% significance level, 156 patients will be required per arm. Allowing for a 30% dropout rate, 224 patients will be required per arm, therefore a total of 448 patients will be recruited. To demonstrate a 17% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and a two-sided significance level, 120 patients will be required per arm. Allowing for a 30% dropout rate, 172 patients will be required per arm, therefore a total of 344 patients will be recruited. To demonstrate a 20% improvement in the response rate for the genotype-guided arm over the non-genotype guided arm with 80% power and a two-sided significance level, 83 patients will be required per arm. Allowing for a 30% dropout rate, 119 patients will be required per arm, therefore a total of 238 patients will be recruited.

## 6.6. Discussion

---

For the primary outcome of the STRONG trial which refers to return to any drinking at 12 months (binary outcome), we have calculated the total sample size

required for four potential designs. The estimated total sample size for each design is given in Table 6.6.

**Table 6.6.** Required total number of patients for four potential designs applied to STRONG trial.

Design	Total sample size for both arms with 30% dropout rate	Total sample size per arm without dropout rate	Difference in response rates
<b>Biomarker-strategy</b>	10356	3624	3% (between genotype-guided arm and non-genotype-guided arm)
	3648	1276	5% (between genotype-guided arm and non-genotype-guided arm)
	852	298	10% (between genotype-guided arm and non-genotype-guided arm)
<b>Marker Stratified</b>	3982	2787	5% (between naltrexone and acamprosate in positive subgroup) and -20% (between naltrexone and acamprosate in negative subgroup)
	1339	937	10% (between naltrexone and acamprosate in positive subgroup) and -15% (between naltrexone and acamprosate in negative subgroup)
	801	561	19.1% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)

	4186	2930	5% (between naltrexone and acamprosate in positive subgroup) and -15% (between naltrexone and acamprosate in negative subgroup)
	4319	3023	5% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)
	1134	794	10% (between naltrexone and acamprosate in positive subgroup) and -20% (between naltrexone and acamprosate in negative subgroup)
	1472	1030	10% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)
	465	325	19.1% (between naltrexone and acamprosate in positive subgroup) and -20% (between naltrexone and acamprosate in negative subgroup)
	669	468	19.1% (between naltrexone and acamprosate in positive subgroup) and -15% (between naltrexone and acamprosate in negative subgroup)
<b>Sequential Subgroup-Specific</b>	18598	13019	5% (between naltrexone and acamprosate in positive subgroup)
	4364	3055	10% (between naltrexone and acamprosate in positive subgroup)

	1014	710	19.1% (between naltrexone and acamprosate in positive subgroup)
<b>Parallel Subgroup- Specific</b>	4821	3375	5% (between naltrexone and acamprosate in positive subgroup) and -20% (between naltrexone and acamprosate in negative subgroup)
	1620	1134	10% (between naltrexone and acamprosate in positive subgroup) and -15% (between naltrexone and acamprosate in negative subgroup)
	971	680	19.1% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)
	5067	3547	5% (between naltrexone and acamprosate in positive subgroup) and -15% (between naltrexone and acamprosate in negative subgroup)
	5230	3661	5% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)
	1374	962	10% (between naltrexone and acamprosate in positive subgroup) and -20% (between naltrexone and acamprosate in negative subgroup)



	1782	1248	10% (between naltrexone and acamprosate in positive subgroup) and -13.2% (between naltrexone and acamprosate in negative subgroup)
	563	394	19.1% (between naltrexone and acamprosate in positive subgroup) and -0.20% (between naltrexone and acamprosate in negative subgroup)
	809	566	19.1% (between naltrexone and acamprosate in positive subgroup) and -0.15% (between naltrexone and acamprosate in negative subgroup)
<b>Reverse Marker- Based strategy</b>	448	156	15% (between genotype-guided arm and reverse-genotype-guided arm)
	334	172	17% (between genotype-guided arm and reverse-genotype-guided arm)
	238	119	20% (between genotype-guided arm and reverse-genotype-guided arm)

In this chapter, we sought to demonstrate a potential procedure for researchers to follow in order to identify the most appropriate type of trial for their particular case. After answering a list of questions that we developed to assist in making this decision and calculating the required sample size, we demonstrate that Reverse Marker-Based strategy design is the optimal design for the STRONG trial when the endpoint is binary. It requires fewer patients than the other three non-adaptive trial designs that have been proposed for the STRONG trial and it is ethically acceptable since both treatments (i.e. naltrexone and acamprosate) are recommended. This work could be extended to time-to-event outcome as can be seen in section 6.6.1. To complement this chapter, in Chapter 7, we will demonstrate the sample size re-

estimation method by incorporating unblinded interim estimates of the effect size into our selected design, due to uncertainty about the true effect size in both the strategy arms.

#### 6.6.1. Non-adaptive design with time-to-event outcome

---

As some participants may be lost to follow-up prior to the 12 month time point at which the primary outcomes is assessed, we propose that trialists consider a time to event outcome ‘time to first drink’ as the primary outcome. This outcome was used either as a primary or secondary outcome in previous trials, e.g. [5, 32].

Assume a time to event outcome will allow observations for the participants to be censored at the time of dropout. In a survival study (see Chapter 3), the sample size in terms of number of events required in the two treatment groups (experimental and control treatment group) assuming 1:1 randomization is given by

$$D = \frac{4(z_{\alpha} + z_{1-\beta})^2}{[\log(HR)]^2} \quad (6.7)$$

where  $HR$  corresponds to the hazard ratio of the two treatment groups and it is equal to the median survival in the experimental group divided by the median survival in the control group [29]. In the Reverse Marker-Based strategy design, instead of having two treatment arms we have two strategy arms (genotype-guided arm and reverse-genotype-guided arm). Thus, based on (6.7) for time-to-event outcomes, assuming 1:1 randomization between genotype-guided arm and reverse-genotype-guided arm, we propose the following equation as an approximation for the required number of events of the Reverse Marker-Based strategy design, i.e.

$$\begin{aligned}
D &= \frac{4(z_a + z_{1-\beta})^2}{[\log(HR)]^2} = \frac{4(z_a + z_{1-\beta})^2}{\left[\log\left(\frac{m_{strategy\ arm}}{m_{reverse\ arm}}\right)\right]^2} \\
&= \frac{4(z_a + z_{1-\beta})^2}{\left[\log\left(\frac{km_{A+} + (1-k)m_{B-}}{km_{B+} + (1-k)m_{A-}}\right)\right]^2},
\end{aligned} \tag{6.8}$$

where  $m_{strategy\ arm}$ ,  $m_{reverse\ arm}$  are the median times to first drink in the genotype-guided arm and reverse genotype-guided arm respectively and the  $m_{A+}$ ,  $m_{B-}$ ,  $m_{B+}$ ,  $m_{A-}$  are the median times to first drink of biomarker-positive patients receiving naltrexone, biomarker-negative patients receiving acamprosate, biomarker-positive patients acamprosate and biomarker-negative patients receiving naltrexone respectively. The median of the strategy arm and reverse arm can follow approximately a weighted sum formula similar to [33].

In 2003 a double-blind, placebo-controlled study comparing and combining naltrexone and acamprosate in relapse prevention of alcoholism was published by Kiefer et al. [32]. In this randomized trial, the 160 participants with alcoholism received naltrexone, acamprosate, combination of both treatments, or placebo for 12 weeks. Survival analyses were performed on the lapse events (first alcohol intake). From the curves of the survival probabilities toward the event “first alcohol intake” for each of the treatment groups in Kiefer et al. [32] we can derive the median time to first drink for patients receiving acamprosate, i.e.  $m_B \approx 4.3$  weeks and the median time to first drink for patients receiving the naltrexone, i.e.  $m_A \approx 6.7$  weeks. For acamprosate, we assume the same median time to first drink for both biomarker-positive and biomarker-negative patients (i.e.  $m_B \approx m_{B+} \approx m_{B-} \approx 4.3$ ) as in the case when assuming a binary outcome. For our optimal design (i.e. Reverse Marker-Based Strategy design), when the outcome is binary, the difference in response rates between the reverse-genotype-guided arm and the genotype-guided arm is estimated approximately 15%. Hence, for the time-to-event outcome, we will vary the effect size (i.e. hazard ratio between the two strategy arms) between 0.75 and 0.85 which

correspond to a moderate hazard ratio in order to calculate the required sample size in terms of number of events. These scenarios of effect sizes should be tested from a panel of expert clinicians who will decide which of these effect sizes can be considered as a minimally clinically significant difference. Based on formula (6.8), 380 events are needed at a two-sided 5% significance level to achieve 80% power when the hazard ratio is 0.75. When the hazard ratio is 0.77, 0.81 and 0.84, we need 460, 708 and 1034 events respectively to achieve 80% power (Table 6.7).

**Table 6.7.** Required total number of events and corresponding hazard ratio of the Reverse Marker-Based strategy design applied to STRONG trial.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
<b>Hazard ratio</b>	0.75	0.77	0.81	0.84
<b>Two-sided significance level (<math>\alpha</math>)</b>	0.05	0.05	0.05	0.05
<b>Power (<math>1 - \beta</math>)</b>	0.8	0.8	0.8	0.8
<b>Total number of events (<math>D</math>) calculated by (7.8)</b>	380	460	708	1034

## 6.7. References

1. Wanless D. Securing Good Health for the Whole Population HMS Treasury Department 2014 [accessed on 10 February 2017]. Available online: <https://www.southampton.gov.uk/moderngov/documents/s19272/prevention-appx%201%20wanless%20summary.pdf>.
2. Department of Health. National Harm Reduction Strategy 2004 [accessed on 10 February 2017]. Available online: <http://www.ave.ee/download/Alcohol%20England.pdf>.

3. Pirmohamed M, Brown C, Owens L, Luke C, Gilmore I, Breckenridge A, et al. The burden of alcohol misuse on an inner-city general hospital. *QJM-AN INTERNATIONAL JOURNAL OF MEDICINE*. 2000; 93(5).
4. Rosner S, Hackl-Herrwerth A, Leucht S, Vecchi S, Srisurapanont M, Soyka M. Opioid antagonists for alcohol dependence. *COCHRANE DATABASE OF SYSTEMATIC REVIEWS*. 2010; (12). doi: 10.1002/14651858.CD001867.pub2.
5. Rosner S, Hackl-Herrwerth A, Leucht S, Leher P, Vecchi S, Soyka M. Acamprosate for alcohol dependence. *COCHRANE DATABASE OF SYSTEMATIC REVIEWS*. 2010; (9). doi: 10.1002/14651858.CD004332.pub2.
6. NICE. Alcohol use disorders 2011 [accessed on 10 February 2017]. Available online: <https://www.nice.org.uk/guidance/cg115>.
7. Anton RF, Randall CL, O'Malley SS, Ciraulo DA, Cisler RA, Zweben A, et al. Combined pharmacotherapies and behavioral interventions for alcohol dependence: The COMBINE study: A randomized controlled trial. *Journal of the American Medical Association*. 2006; 295(17):2003-17. doi: 10.1001/jama.295.17.2003. PubMed PMID: edselc.2-52.0-33646178951.
8. Bouza C, Angeles M, Munoz A, Amate JM. Efficacy and safety of naltrexone and acamprosate in the treatment of alcohol dependence: a systematic review. *Addiction*. 2004; 99(7):811-28.
9. Srisurapanont M, Jarusuraisin N. Naltrexone for the treatment of alcoholism: a meta-analysis of randomized controlled trials. *INTERNATIONAL JOURNAL OF NEUROPSYCHOPHARMACOLOGY*. 2005;8(2).
10. Rubio G, Ponce G, Rodriguez-Jiménez R, Jiménez-Arriero MA, Hoenicka J, Palomo T. Clinical predictors of response to naltrexone in alcoholic patients: who

benefits most from treatment with naltrexone? *Alcohol And Alcoholism* (Oxford, Oxfordshire). 2005; 40(3):227-33. PubMed PMID: 15797885.

11. Anton RF, Oroszi G, O'Malley S, Couper D, Swift R, Pettinati H, et al. An evaluation of mu-opioid receptor (OPRM1) as a predictor of naltrexone response in the treatment of alcohol dependence: results from the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) study. *Archives Of General Psychiatry*. 2008; 65(2):135-44. doi: 10.1001/archpsyc.65.2.135. PubMed PMID: 18250251.

12. Oroszi G, Anton RF, O'Malley S, Swift R, Pettinati H, Couper D, et al. OPRM1 Asn40Asp predicts response to naltrexone treatment: a haplotype-based approach. *Alcoholism: Clinical & Experimental Research*. 2009; 33(3):383-93. doi: 10.1111/j.1530-0277.2008.00846.x.

13. Oslin DW, Berrettini W, Kranzler HR, Pettinati H, Gelernter J, Volpicelli JR, et al. A functional polymorphism of the mu-opioid receptor gene is associated with naltrexone response in alcohol-dependent patients. *NEUROPSYCHOPHARMACOLOGY*. 2003; 28(8).

14. Anton RF. Genetic basis for predicting response to naltrexone in the treatment of alcohol dependence. *Pharmacogenomics*. 2008; 9(6):655-8. doi: 10.2217/14622416.9.6.655.

15. Ray LA, Hutchison KE. A polymorphism of the mu-opioid receptor gene (OPRM1) and sensitivity to the effects of alcohol in humans. *ALCOHOLISM-CLINICAL AND EXPERIMENTAL RESEARCH*. 2004; 28(12):1789-95.

16. Ray LA, Hutchison KE. Effects of naltrexone on alcohol sensitivity and genetic moderators of medication response - A double-blind placebo-controlled study. *ARCHIVES OF GENERAL PSYCHIATRY*. 2007; 64(9):1069-77.

17. Vallender EJ, Rüedi-Bettschen D, Miller GM, Platt DM, Vallender EJ, Rüedi-Bettschen D, et al. A pharmacogenetic model of naltrexone-induced attenuation of alcohol consumption in rhesus monkeys. *Drug & Alcohol Dependence*. 2010; 109(1-3):252-6. doi: 10.1016/j.drugalcdep.2010.01.005.
18. Gelernter J, Zhang H, Cramer J, Rosenheck R, Krystal JH, Gueorguieva R, et al. Opioid receptor gene (OPRM1, OPRK1, and OPRD1) variants and response to naltrexone treatment for alcohol dependence: Results from the VA Cooperative Study. *Alcoholism: Clinical and Experimental Research*. 2007; 31(4):555-63. doi: 10.1111/j.1530-0277.2007.00339.x.
19. Kim S-G, Kim C-M, Choi S-W, Jae Y-M, Lee H-G, Son B-K, et al. A micro opioid receptor gene polymorphism (A118G) and naltrexone treatment response in adherent Korean alcohol-dependent patients. *Psychopharmacology*. 2009; 201(4):611-8. doi: 10.1007/s00213-008-1330-5. PubMed PMID: 18795264.
20. Ray LA, Bujarski S, Chin PF, Miotto K. Pharmacogenetics of naltrexone in asian americans: a randomized placebo-controlled laboratory study. *Neuropsychopharmacology*. 2012; 37(2):445-55. doi: 10.1038/npp.2011.192.
21. Kiefer F, Helwig H, Tarnaske T, Otte C, Jahn H, Wiedemann K. Pharmacological relapse prevention of alcoholism: clinical predictors of outcome. *European Addiction Research*. 2005; 11(2):83-91.
22. Whitworth AB, Fischer F, Lesch OM, Nimmerrichter A, Oberbauer H, Platz T, et al. Comparison of acamprosate and placebo in long-term treatment of alcohol dependence. *Lancet*. 1996; 347 North American Edition (9013):1438-42.
23. Kiefer F, Witt SH, Frank J, Richter A, Treutlein J, Lemenager T, et al. Involvement of the atrial natriuretic peptide transcription factor GATA4 in alcohol dependence,

relapse risk and treatment response to acamprosate. *Pharmacogenomics Journal*. 2011; 11(5):368-74. doi: 10.1038/tpj.2010.51.

24. Ooteman W, Naassila M, Koeter MW, Verheul R, Schippers GM, Houchi H, et al. Predicting the effect of naltrexone and acamprosate in alcohol-dependent patients using genetic indicators. *Addiction Biology*. 2009; 14(3):328-37. doi: 10.1111/j.1369-1600.2009.00159.x.

25. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: design issues. *JNCI: Journal of the National Cancer Institute*. 2010; 102(3):152-60.

26. Khoury MJ, McBride CM, Schully SD, Ioannidis JPA, Feero WG, Janssens A, et al. The Scientific Foundation for Personal Genomics: Recommendations from a National Institutes of Health-Centers for Disease Control and Prevention Multidisciplinary Workshop. *GENETICS IN MEDICINE*. 2009; 11(8).

27. Project\_MATCH\_Research\_Group. Matching Alcoholism Treatments to Client Heterogeneity: Project MATCH posttreatment drinking outcomes. 1997; 58(1):7-29. doi: <http://dx.doi.org/10.15288/jsa.1997.58.7>.

28. Eng KH. Randomized reverse marker strategy design for prospective biomarker validation. *Statistics in Medicine*. 2014; 33(18):3089-99. doi: 10.1002/sim.6146.

29. Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *Journal of Personalized Medicine*. 2017; 7(1). doi: 10.3390/jpm7010001.

30. Baker SG, Kramer BS, Sargent DJ, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. *Discovery medicine*. 2012; 13(70):187-92.



31. Simon R. The use of genomics in clinical trial design. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2008; 14(19):5984-93. doi: 10.1158/1078-0432.CCR-07-4531.
32. Kiefer F, Jahn H, Tarnaske T, Helwig H, Briken P, Holzbach R, et al. Comparing and Combining Naltrexone and Acamprosate in Relapse Prevention of Alcoholism. *Archives of General Psychiatry*. 2003; 60(1). doi: 10.1001/archpsyc.60.1.92.
33. Mandrekar SJ, Sargent DJ. Clinical Trial Designs for Predictive Biomarker Validation: One Size Does Not Fit All. *Journal of biopharmaceutical statistics*. 2009;19(3):530-542. doi:10.1080/10543400902802458.

## Chapter 7. Case study - An adaptive approach

---

### 7.1. Introduction

---

In the case study of the STRONG trial (Chapter 6) we identified the optimal design, the so-called Reverse Marker-Based strategy design. In the current chapter we consider incorporating sample size re-estimation into our chosen design.. Specifically, we consider an adaptive trial design and sample size re-estimation due to uncertainty about the true effect size in both the genotype-guided arm and the reverse-genotype-guided arm. In this adaptive version of the design, we allow early stopping of the trial at the end of the interim analysis due to efficacy of the study in the case where we have obtained a significant p-value showing that the genotype-guided arm is better than the reverse-genotype-guided arm. This ensures reduction in cost and time of the study as there is no need to enrol further patients to prove efficacy. We also allow the study to stop for futility because of non-promising results at the interim stage.

### 7.2. Choosing the type of Sample size re-estimation method

---

When designing a clinical trial, there is often lack of knowledge regarding the design parameters fed into initial sample size calculations, such that there is uncertainty regarding the validity of sample size estimates for a clinical trial design with fixed sample size. In such cases, an adaptive design which enables adjustment (either increase or decrease) of the number of patients to provide desired study power can be adopted, i.e. the so-called sample size re-estimation design [1]. Here, the initial sample size estimate can be adjusted according to the results of an interim analysis where the accumulating data are examined to check whether they provide information that is consistent with the assumptions made in the initial sample size calculations.

For sample size re-estimation several methods have been recommended in the literature, including blinded, unblinded and mixed methods [1]. More precisely, in

the blinded sample size re-estimation approach the results from interim analysis are used without unblinding treatment assignment to provide an updated estimate of the nuisance parameter (e.g. variance of the outcome) based on which the sample size of the trial will be recalculated. Hence, with this approach only the variance will be estimated at the interim analysis stage and the sample size will be reassessed based on that and the initial assumption of treatment difference. In contrast, the unblinded method uses the knowledge of treatment assignment. Hence, at the interim analysis stage, both the treatment effect and variability estimates will be calculated and they will inform the updated sample size. Similar to the blinded method, the treatment difference is not revealed when using the mixed approach and the sample size adjustment will be a combination of a blinded estimate and a futility boundary which will be based on unblinded data (i.e. blinded approach with a futility analysis added). In cases where there is considerable lack of knowledge about the treatment effect and nuisance parameter, the unblinded sample size re-estimation method might be preferred. For this reason, in the STRONG trial we adopt the unblinded method. Two types of the unblinded sample size recalculation are considered: (i) adjustment based on effect-size ratio and (ii) adjustment based on conditional power. Additionally, as we are assuming a two-stage (i.e. one interim analysis is included) adaptive design, we need to allow for the results from each stage to be combined. To do this, we adopt a common method which is based on the inverse-normal stagewise p-value proposed by Lehmacher and Wassmer [2]. More precisely, when this method is used, the inverse standard normal distribution is applied to the p-values at each stage (i.e. the second stage p-value is not using the stage 1 + stage 2 data), with the aim of obtaining a standardized normal statistic. After that, we obtain the combined standardized normal statistic by combining the normal statistics initially obtained through a weighted sum.

### 7.3. Simulation study 1: With the option of early stopping of the trial for efficacy

---

To illustrate the proposed design, a simulation study is performed for both binary and time-to-event outcomes. In this simulation study, we allow the trial to stop only for efficacy.

#### 7.3.1. Calculation of required sample size

---

First, we calculate the required fixed sample size per arm of the Reverse Marker-Based strategy design.

For a binary outcome, equation (3.38) given in Subsection 3.2.4.4. is used for the calculation of the number of patients per arm, while for an event-time outcome, we calculate the required number of events based on the formula in equation (6.8) in Subsection 6.6.1, and then the required number of patients following formulations in Subsection 3.2.2. of Chapter 3 [3]. More precisely, the total number of patients is calculated by

$$N = \frac{D}{Pr(event)},$$

where  $D$  refers to the required number of events.  $Pr(event)$  corresponds to the probability of observing an event and can be defined as

$$Pr(event) = \pi_{ge}Pr_{ge}(event) + \pi_{re}Pr_{re}(event),$$

where subscripts  $ge$  and  $re$  refer to genotype-guided strategy arm and reverse-genotype-guided strategy arm respectively and  $\pi_{ge}$  and  $\pi_{re}$  are the proportions of patients who are randomized to genotype-guided strategy arm and reverse-genotype-guided strategy arm respectively.

Let  $Pr_s(event)$  where  $s$  ( $s = ge$  or  $s = re$ ) be the probability of event in each strategy arm and can be calculated by using the formula given in Kleinbaum and Klein (2012) [4], i.e.

$$Pr_s(event) = 1 - \frac{1}{\left(\frac{\log(2)}{m_s}\right) \times T} \times \left[ e^{-\left(\frac{\log(2)}{m_s}\right) \times \tau} - e^{-\left(\frac{\log(2)}{m_s}\right) \times (T+\tau)} \right],$$

where  $T$  corresponds to the accrual period and  $\tau$  corresponds to the follow-up period.  $m_s$  denotes the median time-to-event of the corresponding treatment strategy arm, and is given by

$$m_s = km_{s+} + (1 - k)m_{s-},$$

where  $k$  denotes the biomarker prevalence and  $m_{s+}$  and  $m_{s-}$  are the median time-to-event in the corresponding treatment strategy arm of biomarker-positive and biomarker-negative subgroups respectively.

### 7.3.2. Target number of patients at the interim stage

---

Next, we calculate the target number of patients to be included in the interim analysis per arm based on a fixed interim fraction  $f$  by

$$N_1 = N_{/arm} \times f.$$

### 7.3.3. Variance of outcome

---

The variance for the two treatment strategy arms, under large sample size assumptions, can be defined for a binary outcome by

$$\sigma = \sqrt{\bar{r}(1 - \bar{r})},$$

[1] where  $\bar{r}$  is the average of the response rates across both trial arms. For an event-time outcome, with uniform patient entry (i.e. time of patient entry for participants is modeled with a Uniform distribution for entry times. The accrual continues until the

assumed accrual period  $T$ . Thus, study entry times are generated from  $U \sim \text{Unif}[T - 1, T]$ , the variance can be defined by

$$\sigma = \bar{r}(1 + u)^{(-1/2)},$$

where

$$u = \frac{\exp[-\bar{r} \times (T + \tau)] \times [1 - \exp(\bar{r} \times T)]}{(T \times \bar{r})},$$

and  $\bar{r}$  in case of time-to-event outcome refers to the average of the hazard rates across both trial arms [1].

#### 7.3.4. Test statistic of the first stage

---

For each simulated iteration, we generate one random number for the genotype-guided arm ( $g_1'$ ) from a normal distribution with mean  $g_1$  and standard deviation  $\sigma/\sqrt{N_1}$  and one random number for the reverse-genotype-guided arm ( $g_4'$ ) with mean  $g_4$  and standard deviation  $\sigma/\sqrt{N_1}$ , independently. When the outcome is binary,  $g_1$  and  $g_4$  refers to the response rate of genotype-guided arm and reverse-genotype-guided arm respectively (see Chapter 3, Subsection 3.2.4.4.). If the outcome is time-to-event,  $g_1$  and  $g_4$  correspond to the hazard rates of genotype-guided arm and reverse-genotype-guided arm respectively. Since

$$g_1' \sim N\left(g_1, \frac{\sigma^2}{N_1}\right)$$

and

$$g_4' \sim N\left(g_4, \frac{\sigma^2}{N_1}\right),$$

then

$$g_1' - g_4' \sim N\left(g_1 - g_4, \frac{2\sigma^2}{N_1}\right).$$

Hence, the test statistic under the null hypothesis  $H_0: g_1 - g_4 = 0$  can be given by

$$T_1 = \frac{(g_1' - g_4') \times \sqrt{\frac{N_1}{2}}}{\sigma}.$$

The corresponding p-value of the observed test statistic  $T_1$  is given by

$$p_1 = Pr(T_1 \geq t_1 | H_0),$$

where  $H_0$  is the null hypothesis of no treatment effect.

### 7.3.5. Stopping boundaries

---

Before conducting a two-stage design, pre-specification of stopping rules and boundaries for efficacy and/or futility are needed. Stopping probabilities (i.e. rejection probabilities), which are calculated based on the stopping boundaries, are essential operating characteristics of adaptive designs. Two types of stopping probabilities exist, the so-called ‘efficacy stopping probability’ which refers to the unconditional probability of rejecting the null hypothesis of no treatment effect, thus the trial stops in order to claim efficacy and the ‘futility stopping probability’ which refers to the unconditional probability of not rejecting the null hypothesis of no treatment effect, thus the trial stops in order to claim futility. In our simulation study, we have considered the two most common types of stopping boundaries which allow early stopping of the trial only for efficacy: (i) O’Brien and Fleming’s boundaries; and (ii) Pocock’s boundaries. Generally, O’Brien-Fleming efficacy boundaries are preferred when the aim is to keep the trial going instead of stopping the trial early due to promising results. On the contrary, with the Pocock efficacy boundaries there is greater chance of stopping the trial early as they sum up the type I error rate earlier. Stopping boundaries for efficacy can be determined by the formula of O’Brien-Fleming and Pocock alpha or  $\alpha$ -spending function which has been defined by Demets and Lan (1994) as “A way of describing the rate at which the total level of significance ( $\alpha$ ) is spent as a continuous function of information fraction and thus induces a

corresponding boundary" [5]. The  $\alpha$ -spending function that approximates the O'Brien-Fleming boundary is [5]:

$$\alpha_{\text{O'Brien-Fleming}}(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{t^*}),$$

and the  $\alpha$ -spending function that approximates Pocock's boundary is [5]:

$$\alpha_{\text{Pocock}}(t^*) = \alpha \ln[1 + (e - 1)t^*],$$

where  $\Phi$  denotes the standard normal cumulative distribution function,  $\alpha$  is the nominal type I error rate (for a one-sided test,  $\alpha = \alpha/2$ ) and  $t^*$  denotes the information fraction, e.g., for a two-stage design, an equally spaced  $t^*$  is equal to 1/2 in the first stage and 2/2 in the second stage. Additionally,  $\alpha_{\text{O'Brien-Fleming}}(t^*)$  and  $\alpha_{\text{Pocock}}(t^*)$  are the  $\alpha$ -spending functions at interim time  $t^*$  for a one-sided test. These  $\alpha$ -spending functions present the cumulative type I error rate of O'Brien-Fleming and Pocock boundaries spent up to the information time  $t^*$ . Consequently, the type I error to spend at the  $o^{\text{th}}$  stage with information time  $t_o^*$  can be found by

$$\alpha_{\text{O'Brien-Fleming}}(t_o^*) - \alpha_{\text{O'Brien-Fleming}}(t_{o-1}^*) \text{ and } \alpha_{\text{Pocock}}(t_o^*) - \alpha(t_{o-1}^*),$$

for O'Brien-Fleming type and Pocock type boundaries respectively.

Based on the aforementioned stopping boundaries, decisions are made during the interim analysis about whether to stop the trial early for efficacy. More precisely, if  $p_1 \leq \alpha_1$ , where  $\alpha_1$  refers either to the type I error at stage 1 calculated by using the O'Brien-Fleming or Pocock method and  $p_1$  is the p-value of the observed value  $t_1$  of the test statistic at stage 1, the trial is stopped for efficacy at stage 1, meaning that the null hypothesis of no treatment effect is rejected. Otherwise, the trial continues to the second stage.



### 7.3.6. Sample size adjustment

---

If the trial has been stopped for efficacy at the first stage of the study, the sample size equals  $N_1$ , however, if the trial continues to the second stage, we recalculate the sample size.

- i) If the adjustment is chosen based on effect-size ratio [1], then the formation of the sample size adjustment can be given by

$$N_{final} = \min \left\{ N_{max}, \max \left[ N_{arm}, \left( \frac{g_1 - g_4}{|g_1' - g_4'| + 0.0000001} \right)^{tu} \times N_{arm} \right] \right\},$$

where  $N_{final}$  is the newly estimated number of patients per group,  $N_{arm}$  is the estimated sample size per group of a non-adaptive design,  $N_{max}$  is a prespecified parameter in our simulations which refers to the chosen upper limit of total sample size per group and  $tu$  is a tuning parameter that is often chosen to be 2 [6]. The small number 0.0000001 is added to avoid numerical overflow if  $g_1' - g_4' = 0$ . In real practice, the allowed maximum sample size  $N_{max}$  will be determined by the sponsors of the trial who will decide a suitable sample size extension.

- ii) If an adjustment based on conditional power with the limit of the maximum number of patients ( $N_{max}$ ) allowed due to cost and time considerations is selected [1], then the formation of the sample size can be given by

$$N_2 = \min \left[ N_{max} - N_1, \frac{2\sigma^2}{(g_1 - g_4)^2} \left( \frac{z_{1-a_2} - w_1 z_{1-p_1}}{\sqrt{1 - w_1^2}} - z_{1-cP} \right)^2 \right],$$

where  $a_2$  is the final efficacy stopping boundary,  $cP$  is the target conditional power (pre-specified parameter in our simulations which refers to an approach that quantifies the probability of rejecting the null hypothesis of no effect once some data are available) which might be determined by funding bodies in real practice,  $p_1$  is the p-value of the observed test statistic  $t_1$  at stage 1 and  $z_{1-a_2} = \Phi^{-1}(1 - a_2)$ ,  $z_{1-cP} = \Phi^{-1}(1 - cP)$ ,  $z_{1-p_1} = \Phi^{-1}(1 - p_1)$ , where  $\Phi$  denotes the

standard normal cumulative distribution function. Information from both stages is combined using the inverse normal combination test where  $w_1$  and  $w_2$  are pre-specified weights satisfying  $w_1^2 + w_2^2 = 1$  [2]. In our simulation, we fix  $w_1 = w_2 = 1/\sqrt{2}$ .

### 7.3.7. Test statistic of the second stage

---

A similar process used for the determination of the first stage test statistic is followed for the determination of the second stage test statistic (see Section 7.3.4.). The difference is that test statistic of the second stage will be based on the adjustment of the sample size. More precisely, after the recalculation of the sample size,

- i) when the adjustment is based on effect-size ratio method, we generate one random number for the genotype-guided arm ( $g_1''$ ) from normal distribution with mean  $g_1$  and the new standard deviation  $\sigma/\sqrt{N_{final} - N_1}$  and one random number for the reverse-genotype-guided arm ( $g_4''$ ) with mean  $g_4$  and the new standard deviation  $\sigma/\sqrt{N_{final} - N_1}$ . Therefore, the test statistic of the second stage is given by

$$T_2 = \frac{(g_1'' - g_4'') \times \sqrt{\frac{N_{final} - N_1}{2}}}{\sigma}.$$

- ii) when the sample size adjustment is based on the conditional power method, we generate one random number for the genotype-guided arm ( $g_1''$ ) from normal distribution with mean  $g_1$  and the new standard deviation  $\sigma/\sqrt{N_2}$  and one random number for the reverse-genotype-guided arm ( $g_4'$ ) with mean  $g_4$  and the new standard deviation  $\sigma/\sqrt{N_2}$ . The test statistic of the second stage is given by

$$T_2 = \frac{(g_1'' - g_4'') \times \sqrt{\frac{N_2}{2}}}{\sigma}.$$

### 7.3.8. Final test statistic

---

The test statistic of the final analysis (inverse normal combination test) is based on the weighted sum of test statistics of each stage and can be given by

$$Z_2 = w_1 t_1 + w_2 t_2 = \frac{t_1 + t_2}{\sqrt{2}},$$

according to Lehmacher and Wassmer (1999) who suggested the use of equal weights, i.e.  $1/\sqrt{s}$ , where  $s$  denotes the number of stages of the design [2]. The corresponding  $p$ -value to the final test statistic  $Z_2$  is given by

$$p_2 = Pr(Z_2 \geq z_2 | H_0),$$

where  $H_0$  is the null hypothesis of no treatment effect. If  $p_2 \leq a_2$ , where  $a_2$  refers to the type I error at stage 2 calculated by using either the O'Brien-Fleming or Pocock method, we get the total study power.

### 7.3.9. Simulation parameters

---

In our simulation study we consider both types of boundaries to preserve the overall type I error rate for effectiveness at the one-sided 0.025 level. More precisely, we have used the O'Brien-Fleming's efficacy stopping boundaries 0.001525323 and 0.02347468 for stage 1 and stage 2 respectively and the Pocock efficacy stopping boundaries 0.01550286 and 0.009497137 for stage 1 and stage 2 respectively calculated by the package GroupSeq in R statistical software. The weights for combining both stages are also pre-specified as  $w_1 = w_2 = 1/\sqrt{2}$  according to Lehmacher and Wassmer (1999) to ensure control of the type I error rate [2]. In addition, we assume that a reasonable sample size extension is no more than 100 patients. Hence, we allow recruitment of an additional 100 patients ( $N_{max} = N_{arm} + 100$ ) as an upper limit of the allowed sample size. The type II error rate,  $\beta = 0.2$ , was assumed for the calculation of the fixed number of events and patients required in a standard non-adaptive design. Three cases of information fraction were explored, i.e. (i)  $f = 0.25$ , (ii)  $f = 0.50$ , (iii)  $f = 0.75$  in order to investigate a range from low to high percentage.

We performed 1000000 simulated iterations for binary and time-to-event outcome using both the effect size method and the conditional power method.

In the simulation study for binary outcome, the response rate in the genotype-guided arm is assumed to be  $g_1 = 0.72$  and in the reverse genotype-guided arm is  $g_4 = 0.57$  based on the non-adaptive calculations of the Reverse Marker-Based strategy design in Chapter 6. These response rates result in 15% difference in response rates between the two strategy arms.

The time-to-event outcome is based on the study by Kiefer et al. [7] in which participants with alcoholism were randomized between naltrexone, acamprosate, combination of both treatments, or placebo with 12 weeks follow-up period and the accrual period 2 years (i.e 104.28 weeks). In our case, the accrual time and total study period were fixed at 104.28 weeks and 116.28 weeks respectively which correspond to 12 weeks of follow-up.

In the simulation study for time-to-event outcome, four scenarios of median time-to-event for biomarker-positive and biomarker-negative patients receiving the experimental treatment were considered as in Chapter 6, i.e. (i)  $m_{A+} = 6$ ,  $m_{A-} = 6.7$ , (ii)  $m_{A+} = 6.7$ ,  $m_{A-} = 6$ , (iii)  $m_{A+} = 6.6$ ,  $m_{A-} = 6.3$  and (iv)  $m_{A+} = 6.3$ ,  $m_{A-} = 6.6$ . For biomarker-negative and biomarker-positive patients receiving the control treatment, the median time-to-event was 4.3 weeks ( $m_{B-} = m_{B+} = 4.3$ ). The corresponding hazard ratios for the four scenarios of median survival times are 0.746, 0.845, 0.807 and 0.765 respectively calculated by using the divisor of formula (6.8) in Chapter 6. The hazard rate in the genotype-guided arm is given by

$$g_1 = \frac{\ln 2}{m_{strategy\ arm}},$$

where

$$m_{strategy\ arm} = km_{A+} + (1 - k)m_{B-},$$

and the hazard rate in the reverse-genotype-guided arm is given by

$$g_4 = \frac{\ln 2}{m_{reverse\ arm}},$$

where

$$m_{reverse} = km_{B+} + (1 - k)m_{A-}.$$

Simulation parameters for binary and time-to-event outcomes are summarized in Table 7.1.

**Table 7.1.** Summary of simulation parameters for both binary and time-to-event outcomes.

	Binary outcome	Time-to-event outcome	Selection
Efficacy stopping boundary- Stage 1 (scenario 1)	0.001525323	similar to binary outcome	O'Brien-Fleming's boundary
Efficacy stopping boundary- Stage 1 (scenario 2)	0.01550286	similar to binary outcome	Pocock boundary
Efficacy stopping boundary- Stage 2 (scenario 1)	0.02347468	similar to binary outcome	O'Brien-Fleming's boundary
Efficacy stopping boundary- Stage 2 (scenario 2)	0.009497137	similar to binary outcome	Pocock boundary
Weights for combining Stage 1 and Stage 2	$w_1 = w_2 = 1/\sqrt{2}$	similar to binary outcome	Equal weights for simplicity
Upper limit of the allowed sample size	$N_{max} = N_{arm} + 100$	similar to binary outcome	100 patients is assumed as a reasonable sample size extension

			for our trial. Hence, we allow recruitment of an additional 100 patients.
Type II error rate	$\beta = 0.2$	similar to binary outcome	80% power needed for the calculation of the fixed number of events and patients required in a standard non-adaptive design.
Information fraction	(i) $f = 0.25$ , (ii) $f = 0.50$ , (iii) $f = 0.75$	similar to binary outcome	Three cases of information fraction were explored in order to investigate a range from low to high percentage.
Median time-to-event for patients receiving the experimental treatment	Not applicable	(i) $m_{A+} = 6, m_{A-} = 6.7$ , (ii) $m_{A+} = 6.7, m_{A-} = 6$ , (iii) $m_{A+} = 6.6, m_{A-} = 6.3$	Four scenarios of median time-to-event for biomarker-positive and biomarker-negative patients receiving the experimental were considered based on Chapter 6

	(iv) $m_{A+} = 6.3, m_{A-} = 6.6$		
Median time-to-event for patients receiving the control treatment	Not applicable	$m_{B-} = m_{B+} = 4.3$	Median time-to-event for biomarker-positive and biomarker-negative patients receiving the control treatment was based on Chapter 6
Response rate/Hazard rate in genotype-guided arm	$g_1 = 0.72$	$g_1 = \frac{\ln 2}{m_{strategy}}$ <p>where</p> $m_{strategy\ arm} = km_{A+} + (1 - k)m_{B-}$	Calculations based on Chapter 6
Response rate/Hazard rate in reverse-genotype-guided arm	$g_4 = 0.57$	$g_4 = \frac{\ln 2}{m_{reverse\ arm}}$ <p>where</p> $m_{reverse\ arm} = km_{B+} + (1 - k)m_{A-}$	Calculations based on Chapter 6



#### 7.4. Simulation study 2: With the option of early stopping of the trial for efficacy and futility

---

Apart from the efficacy stopping probability, another important design change is the early cessation of the trial due to futility of treatment effect in case that the results are not promising enough at the end of the interim analysis. If the trial claims futility, it does not necessarily mean that the treatment of interest is ineffective, but it shows that even though the study was initially powered at a sufficient high percentage (e.g. 80%), at the interim stage, the revised accumulating information results in much lower power. Therefore, it can be shown that the study will not be plausible as a very large sample size would be required to reach the desired power.

We extend simulation study 1 by allowing the trial to stop early for futility at the end of the interim stage. According to Lan and Demets (2009) [8], one suggestion of futility boundary could be the use of a beta ( $\beta$ ) spending function. More specifically,  $\alpha$  (type I error) can be replaced with  $\beta$  (type II error) in the formula of Pocock type spending function  $\alpha_{\text{Pocock}}(t^*)$ . Therefore, we can have the following  $\beta$ -spending function,

$$\beta_{\text{Pocock}}(t^*) = \beta \ln[1 + (e - 1)t^*].$$

Decisions are made during the interim analysis about whether to stop the trial early for efficacy and futility or to continue to the second stage of the study. More precisely, if  $p_1 > \beta_1$ , where  $\beta_1$  refers to the futility stopping boundary, the trial is stopped for futility at stage 1, meaning that the null hypothesis of no treatment effect is accepted. If  $p_1 \leq \alpha_1$ , where  $\alpha_1$  refers to the type I error at stage 1 calculated by using the O'Brien-Fleming or Pocock method, the trial is stopped for efficacy at stage 1, meaning that the null hypothesis of no treatment effect is rejected. Otherwise, if  $\alpha_1 < p_1 \leq \beta_1$ , the trial continues to the second stage. In our simulation study, the nominal level of power was set at 80%, thus, the type II error rate (i.e.  $\beta$ ) corresponds to 0.2. Therefore, the futility stopping boundary used for our two-stage design is equal to 0.124. The simulation study 1, which allows the trial to stop early only due to efficacy,

follows the same decision process, however,  $\beta_1$  is equal to 1 meaning that we do not allow the study to stop early for futility.

## 7.5. Simulation results

---

Results from Simulation study 1 and Simulation study 2 for both binary and time-to-event outcome are described below.

Graphical summaries are presented in Figures 7.1 to 7.70. The numerical summaries are presented in Appendix D, Tables D.1.1-D.1.4 and D.2.1-D.2.4 for binary outcome from Simulation study 1 and Simulation study 2 respectively. The fixed sample size, difference in response rates between the two strategy arms, average adaptive sample size, efficacy stopping probability and rejection probability are presented. The rejection probability of the null hypothesis corresponds to the type I error rate whereas the rejection probability of the alternative hypothesis corresponds to the total power of the study. Results from Simulation study 1 and 2 for time-to-event outcome are summarized in Appendix D, in Tables D.1.5-D.1.8 and D.2.5-D.2.8 respectively. The fixed number of events and number of patients for the non-adaptive design and the corresponding hazard ratio are presented. Additionally, the average sample size, efficacy and futility stopping probability and rejection probability of the study under the null hypothesis and alternative hypothesis in each of the three cases of information fraction are shown.

### 7.5.1. Binary Outcome

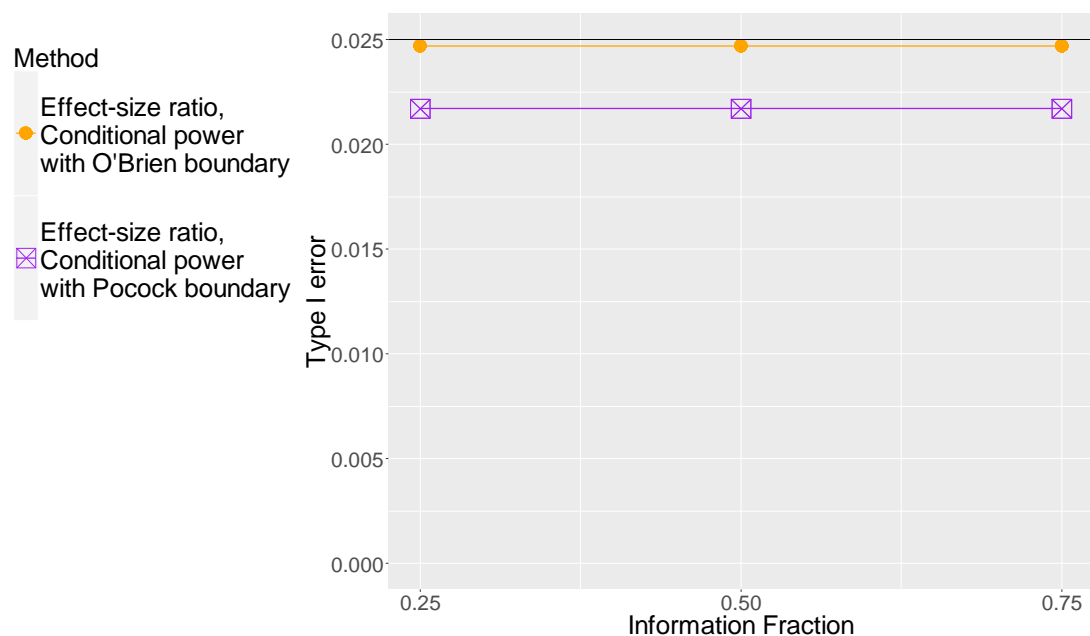
---

#### 7.5.1.1. Control of type I error rate

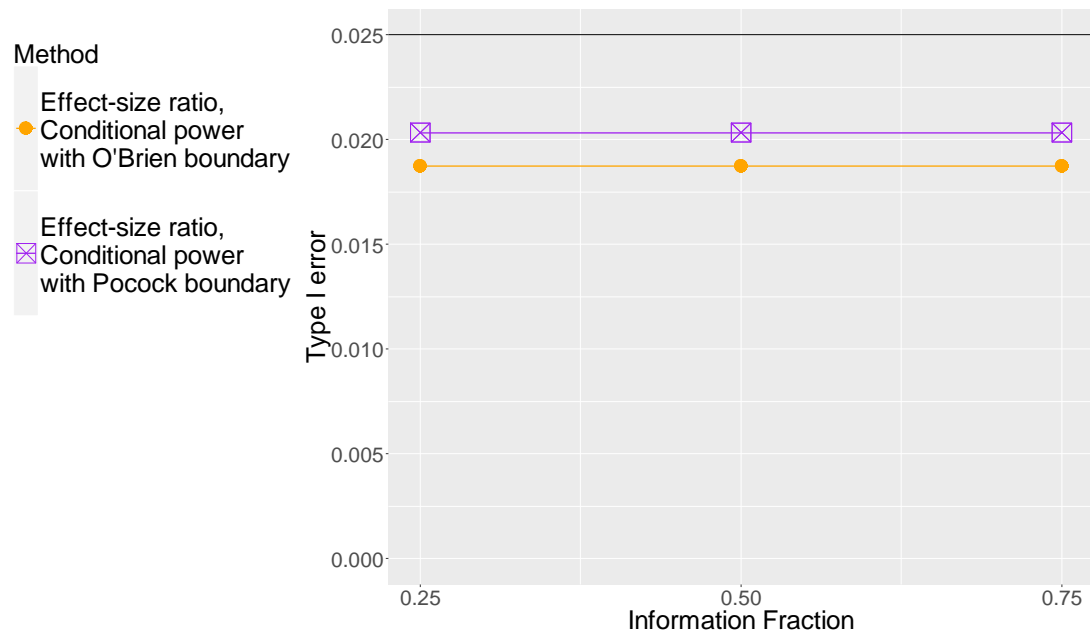
---

In both simulation studies, the type I error rate is well controlled ( $<0.025$ ) when either the effect-size ratio method or the conditional power method for sample size adjustment is used with O'Brien-Fleming and Pocock decision boundaries. More precisely, in Simulation study 1 (i.e. only efficacy stopping of the trial is allowed), the type I error rate is equal to 0.024706 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.1.1-D.1.2) in all scenarios of information fraction

for which the O'Brien-Fleming decision boundaries are applied and it is equal to 0.021698 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.1.3-D.1.4) for Pocock decision boundaries (Figure 7.1). In Simulation study 2 (i.e. both efficacy and futility stopping of the trial are allowed), the type I error rate is equal to 0.018734 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.2.1-D.2.2) in all scenarios for which the O'Brien-Fleming efficacy boundaries are applied and it is equal to 0.020321 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.2.3-D.2.4) for Pocock efficacy boundaries (Figure 7.2). All resulting type I error values shown in Figures 7.1 and 7.2 are under the 0.025 level, indicating that the type I error rate is well controlled.



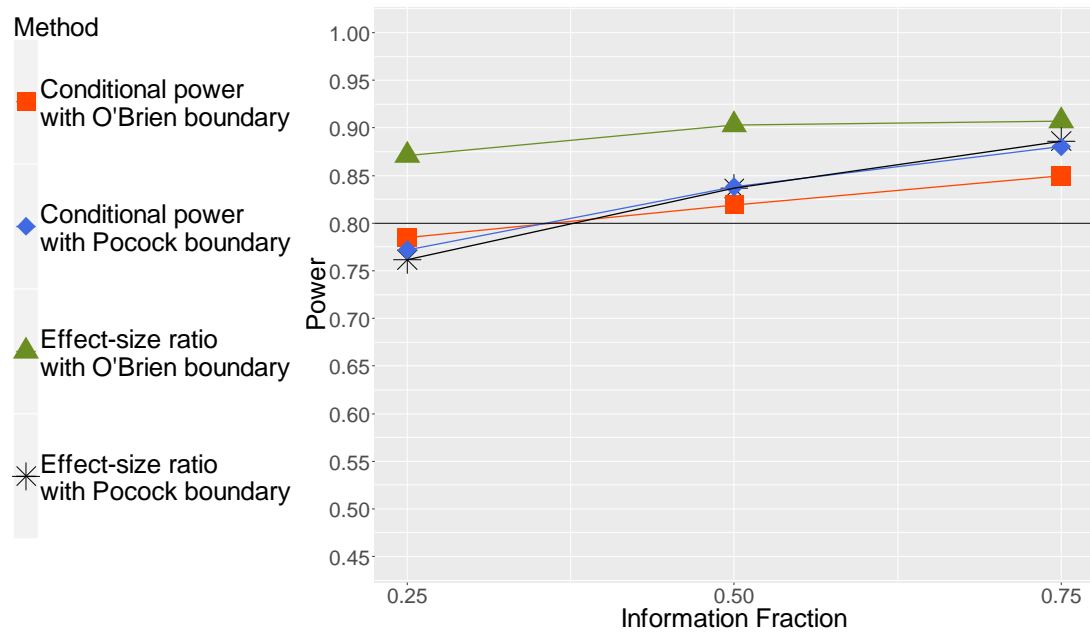
**Figure 7.1.** Type I error versus the information fraction of Simulation study 1. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).



**Figure 7.2.** Type I error versus the information fraction of Simulation study 2. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

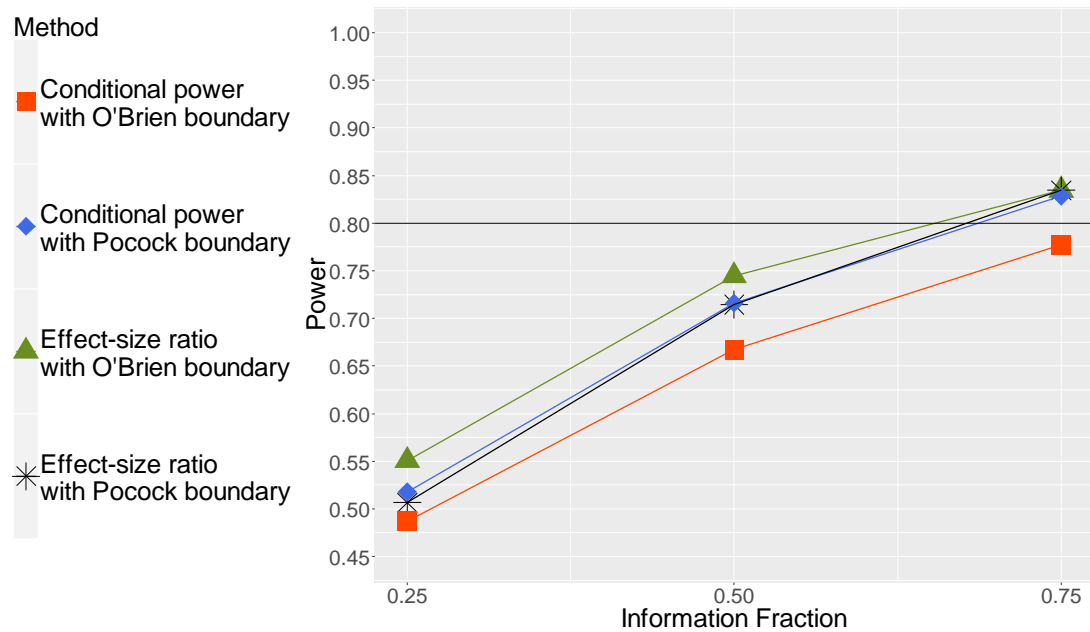
#### 7.5.1.2. Power of the study

Figure 7.3 from Simulation study 1, under the alternative hypothesis, when the effect-size ratio method for sample size adjustment is applied, shows that we can achieve greater power with the O'Brien-Fleming decision boundaries as compared to the Pocock's type boundaries. The conditional power method results in greater power with the O'Brien-Fleming decision boundaries as compared to Pocock boundaries only when the information fraction is 25%. For 50% and 75% information fraction, it seems that Pocock boundaries yield greater power as compared to the O'Brien-Fleming efficacy boundaries. It can also be seen that for 50% and 75% of information fraction, the power is improved compared to the nominal level of power in the non-adaptive design as it exceeds 80% in both types of sample size adjustment and both types of stopping boundaries. At 25% information fraction, the power exceeds the nominal level (i.e. 80%) only when the effect-size ratio method with the O'Brien-Fleming decision boundaries is applied.



**Figure 7.3.** Power versus the information fraction under the alternative hypothesis of Simulation study 1. The horizontal line represents the power of the non-adaptive design (i.e. 80%).

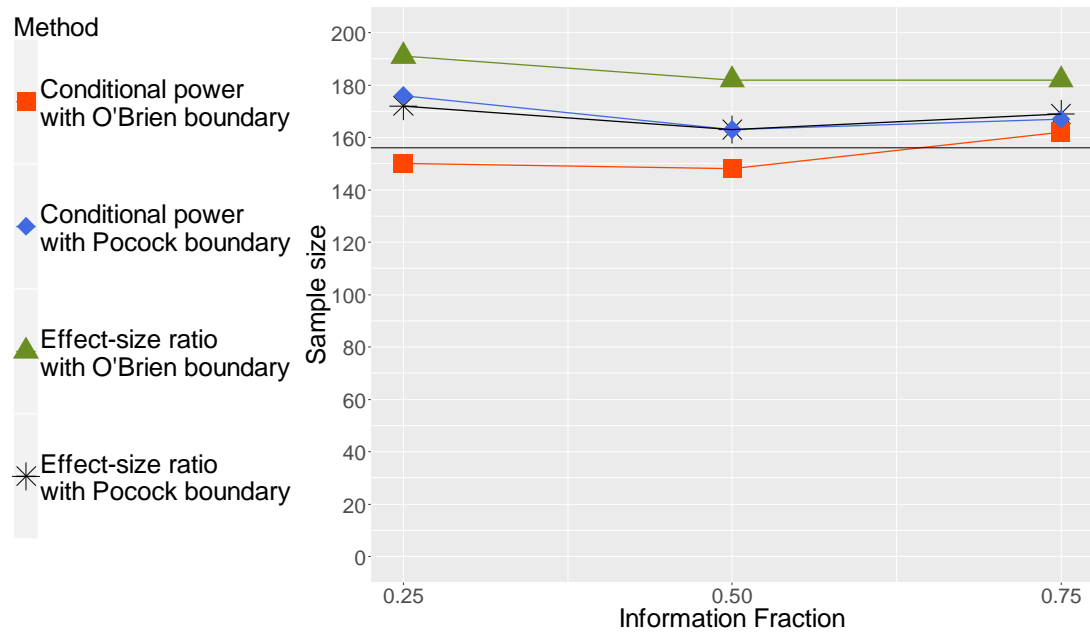
Figure 7.4 from Simulation study 2, under the alternative hypothesis, shows that the power exceeds the nominal level (i.e. 80%) only at 75% information fraction when all types of sample size adjustment methods and stopping boundaries are used apart from the combination of the conditional power method with the O'Brien-Fleming decision boundaries. Among the different sample size adjustment methods and stopping boundaries, the effect-size ratio method with the O'Brien-Fleming decision boundaries results in the highest power and the conditional power method with the same type of stopping boundaries results in the lowest power.



**Figure 7.4.** Power versus the information fraction under the alternative hypothesis of Simulation study 2. The horizontal line represents the power of the non-adaptive design (i.e. 80%).

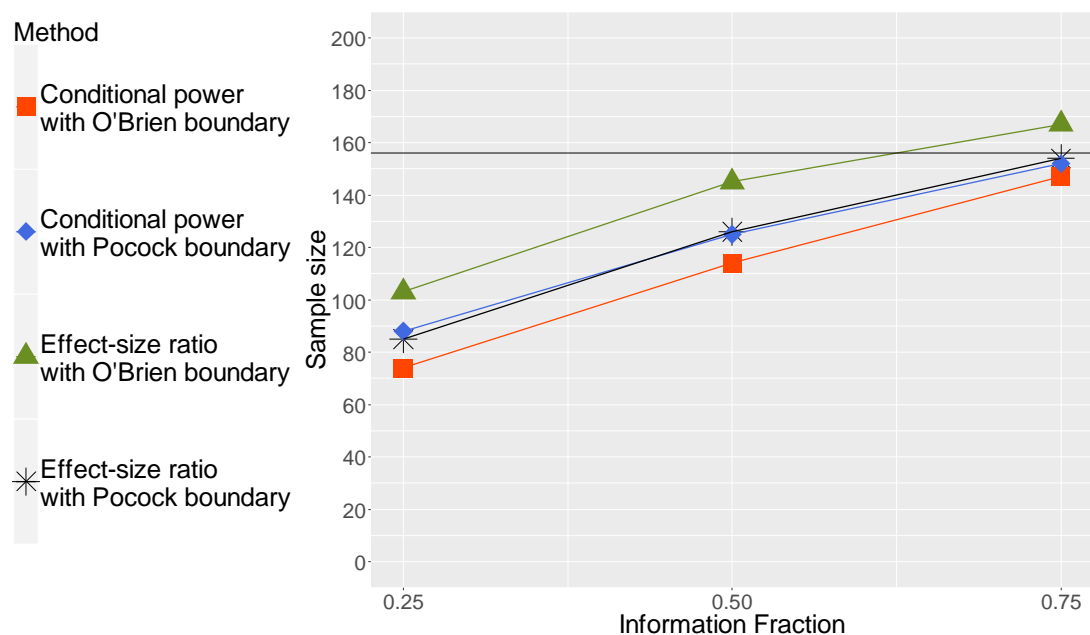
#### 7.5.1.3. Sample size of the study

Figure 7.5 from Simulation study 1, under the alternative hypothesis, shows gain in efficiency (i.e. smaller sample size compared to that of the non-adaptive design) only at 25% and 50% information fraction when the conditional power method with the O'Brien-Fleming boundaries is applied. In all other cases, there is loss of efficiency as the sample size is increased compared to the number of patients required for the non-adaptive design.



**Figure 7.5.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1. The horizontal line represents the sample size of the non-adaptive design.

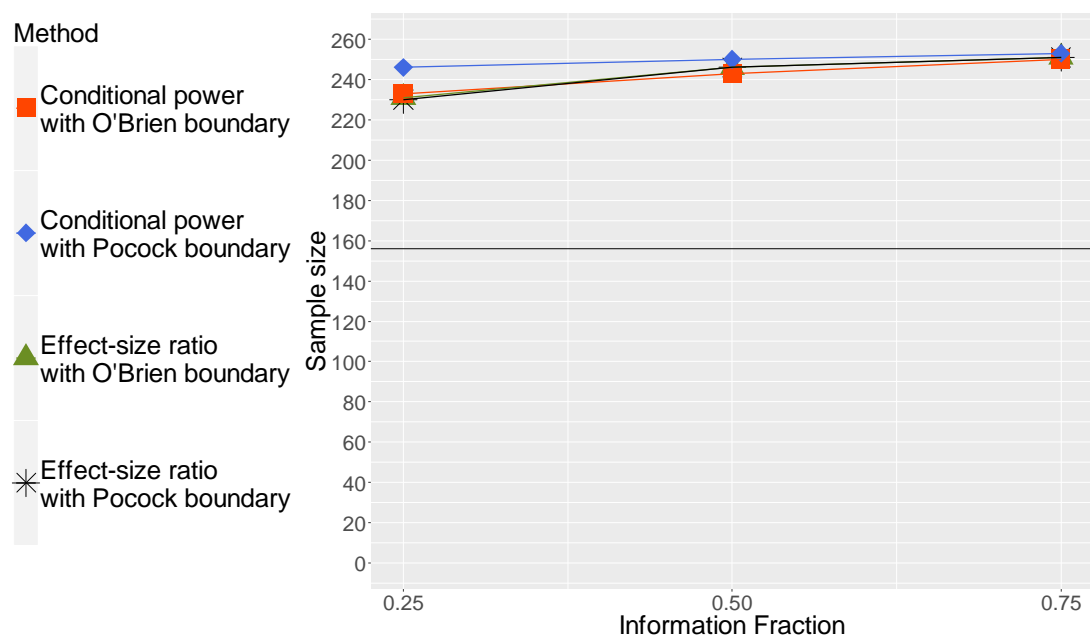
However, this is changed when introducing both efficacy and futility stopping in the trial. It can be seen in Figure 7.6 (Simulation study 2) that the sample size is decreased when using both types of sample size adjustment and stopping boundaries apart from 75% information fraction under effect-size ratio method and O'Brien-Fleming decision boundaries.



**Figure 7.6.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 2. The horizontal line represents the sample size of the non-adaptive design.

Across the different methods for the recalculation of the sample size and stopping boundaries, in both Figures 7.5 and 7.6, the smallest sample size is achieved when the conditional power method with O'Brien-Fleming boundaries is applied and the largest sample size is achieved with the effect-size ratio method and the same type of stopping boundaries.

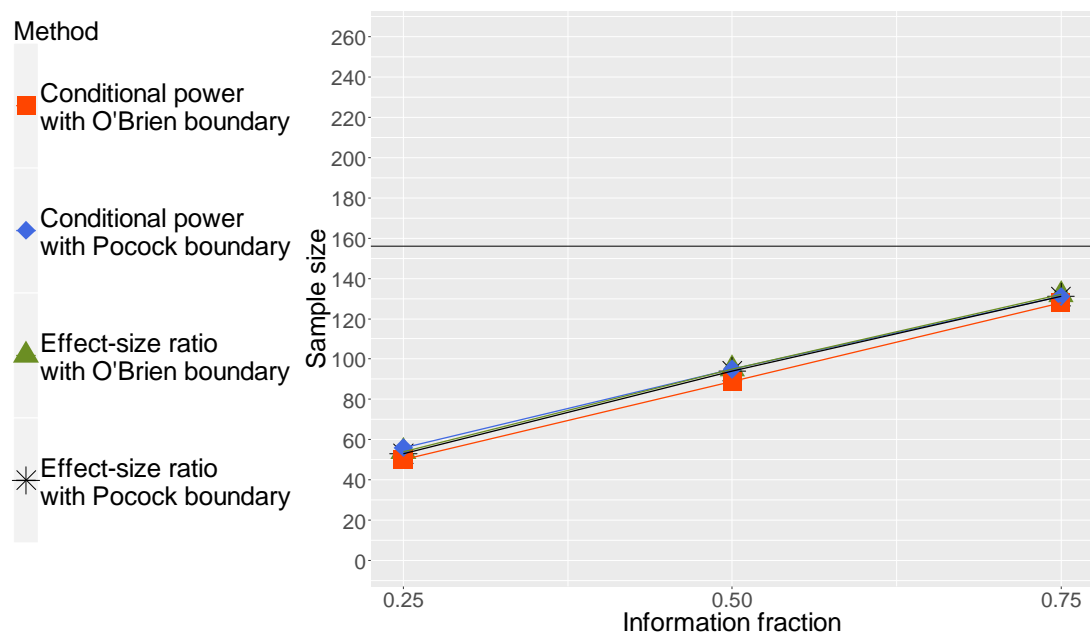
From Simulation study 1, under the null hypothesis, all values are above the horizontal line in Figure 7.7 indicating the increase of the sample size compared to the sample size of the non-adaptive design. The largest increase of the sample size as compared to the number of patients required for the non-adaptive approach corresponds to 75% information fraction. Specifically, with the 75% information fraction, 156 patients per arm required for the non-adaptive approach, and we achieved 250 patients per arm with the conditional power method and O'Brien-Fleming decision boundaries, 251 patients per arm with the effect-size ratio method and the same type of boundaries, 253 patients per arm with the conditional power method and Pocock boundaries and 251 patients per arm with the effect-size ratio method and the same type of stopping boundaries. These numbers are very close to the allowed maximum sample size ( $N_{max}$ ).



**Figure 7.7.** Sample size versus the information fraction under the null hypothesis of Simulation study 1. The horizontal line represents the sample size of the non-adaptive design.



However, this situation changes when introducing the futility stopping in Simulation study 2. As shown in Figure 7.8, all values are under the horizontal line indicating a decrease in sample size compared to a fixed design. Figure 7.8 shows that in both methods of sample size recalculation and both types of efficacy boundaries the number of patients increases when the information fraction increases. The largest increase of the sample size as compared to the number of patients required for the non-adaptive approach corresponds to 75% information fraction. More precisely, with the 75% information fraction, 156 patients per arm were required for the non-adaptive approach, and we would require 128 patients per arm with the conditional power method and O'Brien-Fleming efficacy boundaries, 132 patients per arm with the effect-size ratio method and the same type of stopping boundaries and 131 patients when the conditional power method and the effect-size ratio method with Pocock efficacy boundaries are used.

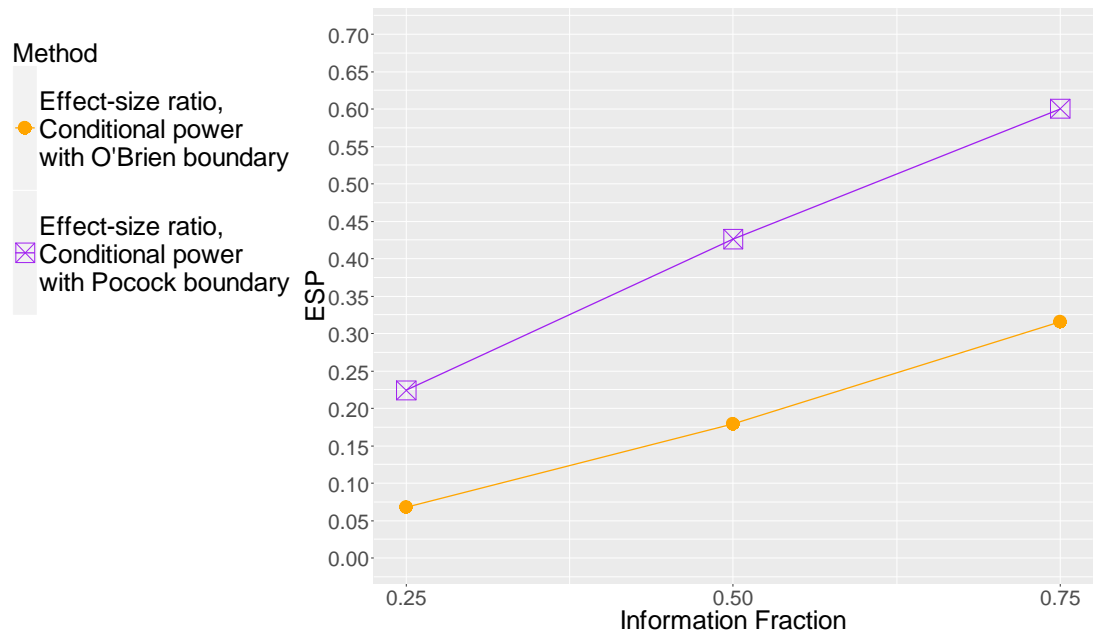


**Figure 7.8.** Sample size versus the information fraction under the null hypothesis of Simulation study 2. The horizontal line represents the sample size of the non-adaptive design.

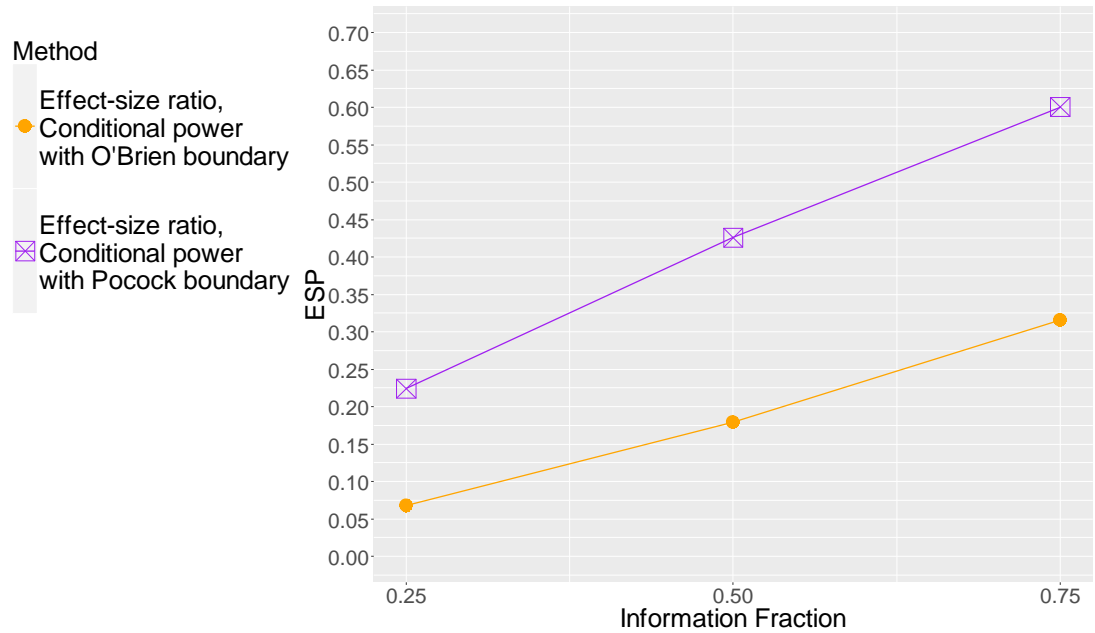
#### 7.5.1.4. Efficacy stopping probability

In both simulation studies, under the alternative hypothesis, the efficacy stopping probability depends only on the type of stopping boundaries. In can be seen in Figures 7.9 and 7.10 that the efficacy stopping probability increases with the

increase of information fraction. Consequently, the largest value is obtained at 75% information fraction. Additionally, the efficacy stopping probability is always greater with the Pocock decision boundaries compared to the probabilities obtained with the O'Brien-Fleming decision boundaries.

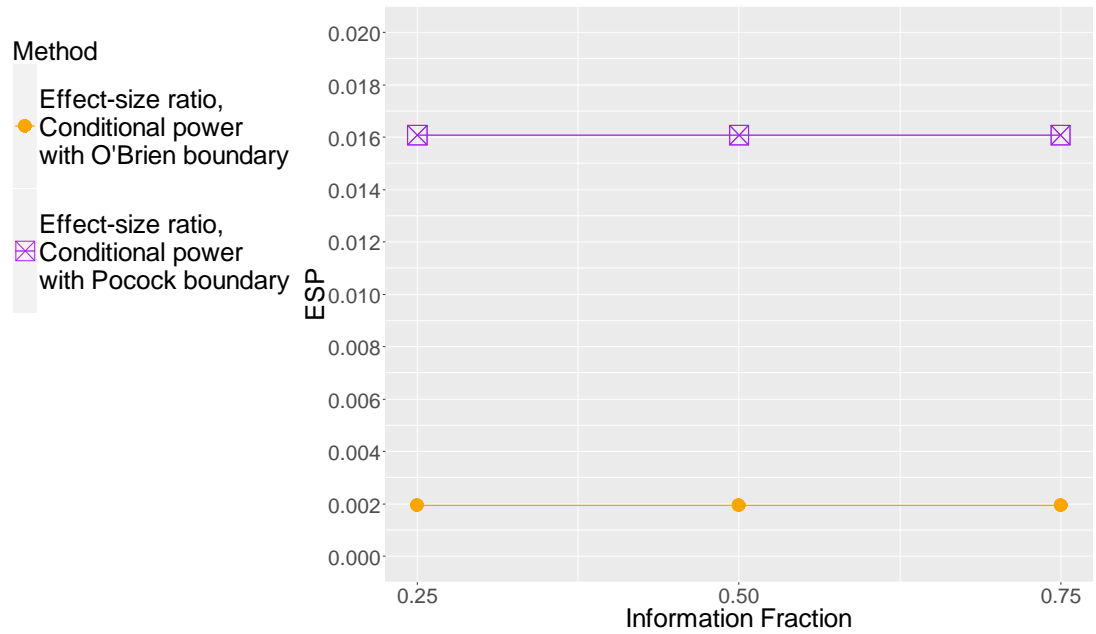


**Figure 7.9.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1.

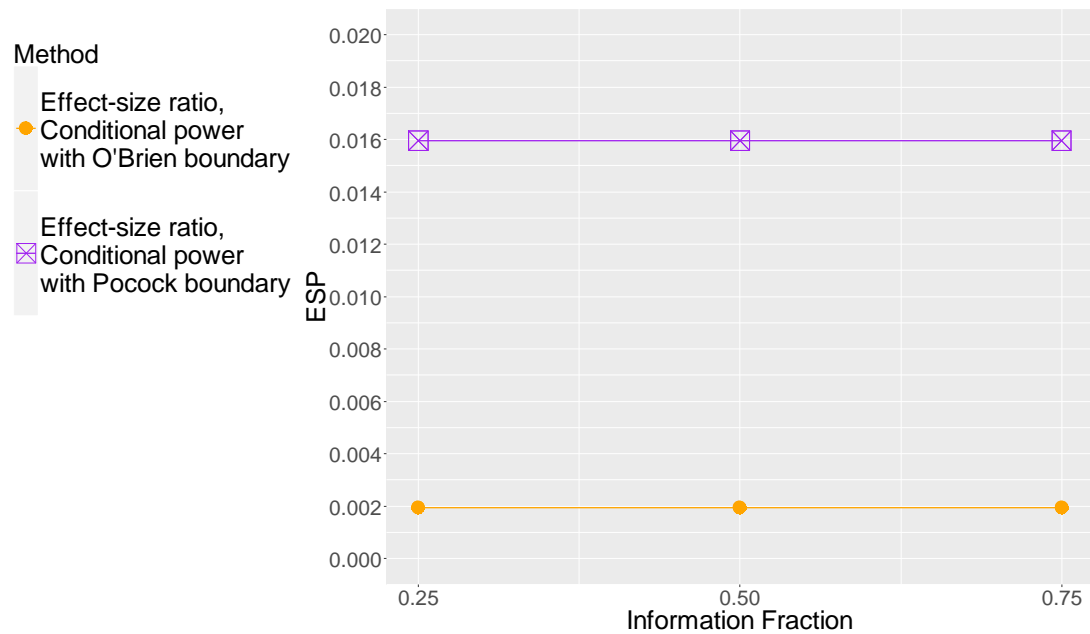


**Figure 7.10.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2.

Similarly to the alternative hypothesis, in both simulation studies under the null hypothesis, the efficacy stopping probabilities depend only on the efficacy stopping boundaries and not on the sample size adjustment methods. Figures 7.11, 7.12 show that the efficacy stopping probabilities remain the same across the different percentages of information fraction for each type of stopping boundaries. Higher values are achieved when the Pocock stopping boundaries are used compared to the O’Brien-Fleming efficacy boundaries.



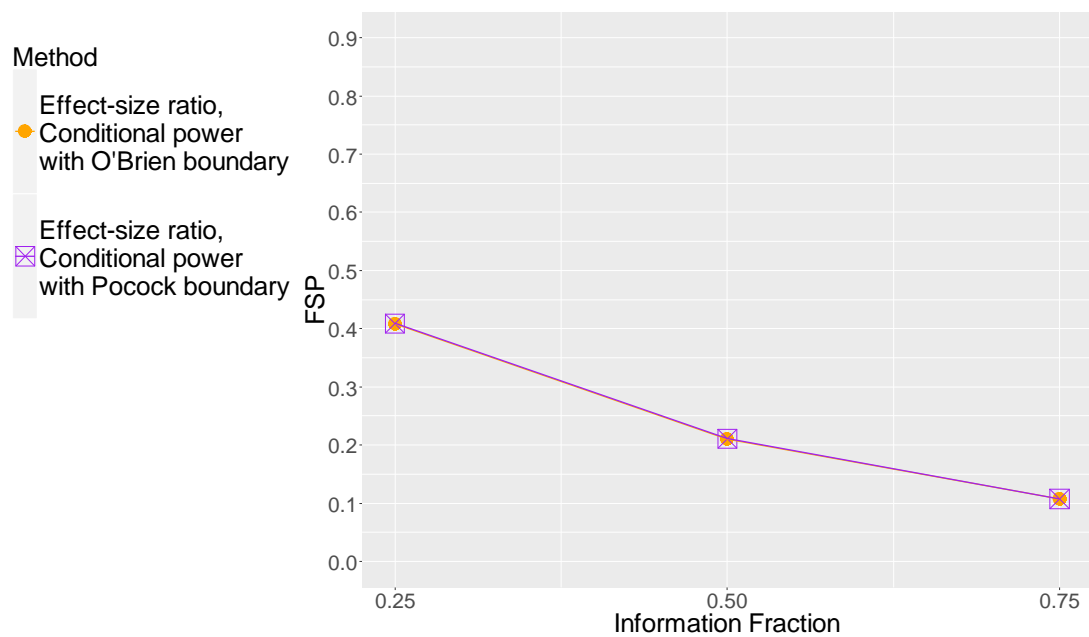
**Figure 7.11.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1.



**Figure 7.12.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2.

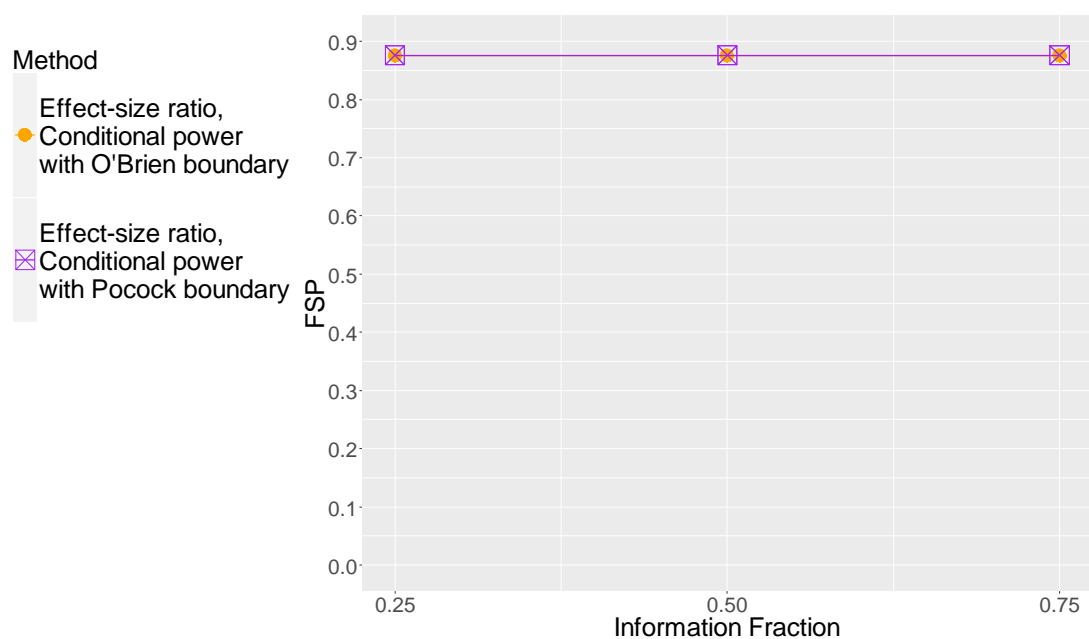
#### 7.5.1.5. Futility stopping probability

In Simulation study 2, the futility stopping probability depends only on the type of stopping boundaries and not on the method of sample size recalculation. As shown in Figure 7.13, futility stopping probability (FSP) decreases with the increase of information fraction under the alternative hypothesis and it depends only on the type of stopping boundaries and not on the method of sample size recalculation. Hence, the smallest value of futility probability is obtained at 75% information fraction. The application of Pocock stopping boundaries results in slightly higher futility stopping probabilities compared to the use of the O'Brien-Fleming efficacy boundaries.



**Figure 7.13.** Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2.

Similarly to the alternative hypothesis, the futility stopping probabilities under the null hypothesis of Simulation study 2, depend only on the efficacy stopping boundaries and not on the sample size adjustment methods. Figure 7.14 shows that the futility stopping probabilities remain the same across the different percentages of information fraction for each type of stopping boundaries. Similar values are achieved with each type of stopping boundaries.



**Figure 7.14.** Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2.

## 7.5.2. Time-to-event Outcome

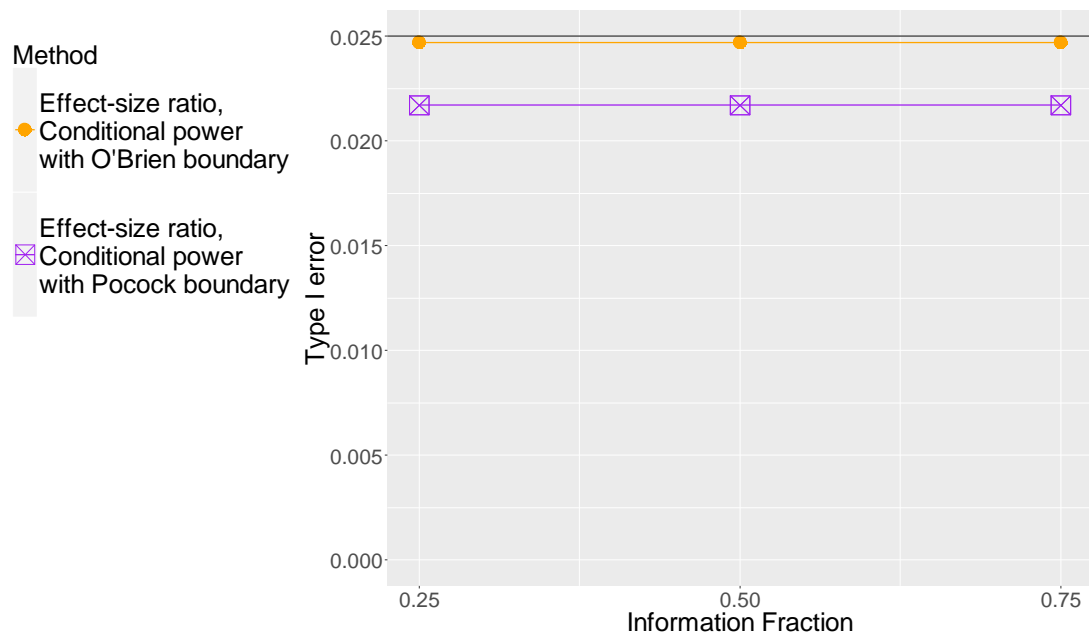
---

### 7.5.2.1. Control of type I error rate

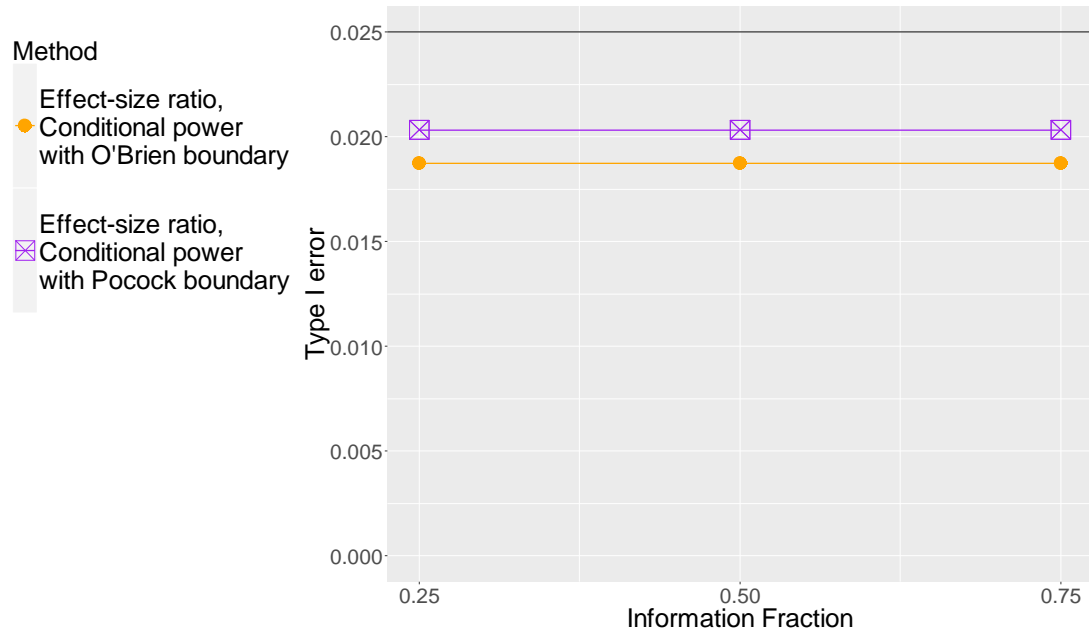
---

Similarly to the binary outcome, it can be seen from our results that the type I error rate is well controlled ( $<0.025$ ) also in the case of survival outcome. The type I error rate in each simulation study is the same with that in binary outcome and all the different scenarios of hazard ratio have the same type I error rate. More precisely, in Simulation study 1, the type I error rate is equal to 0.024706 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.1.5-D.1.6) in all scenarios of information fraction and hazard ratios for which the O'Brien-Fleming decision boundaries are applied and it is equal to 0.021698 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.1.7-D.1.8) for Pocock decision boundaries. In Simulation study 2, the type I error rate is equal to 0.018734 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.2.5-D.2.6) in all scenarios of information fraction and hazard ratio for which the O'Brien-Fleming efficacy boundaries are applied and it is equal to 0.020321 (i.e. rejection probability under the null hypothesis in Appendix D, Tables D.2.7-D.2.8) for Pocock efficacy boundaries. The results of type I error rate versus the information fraction in both simulation studies for each scenario of hazard ratio are shown in Figures 7.15-7.22 below. In all cases, the type I error rate is well controlled as all resulting values are under 0.025 level.

### First scenario of hazard ratio (i.e. 0.746)

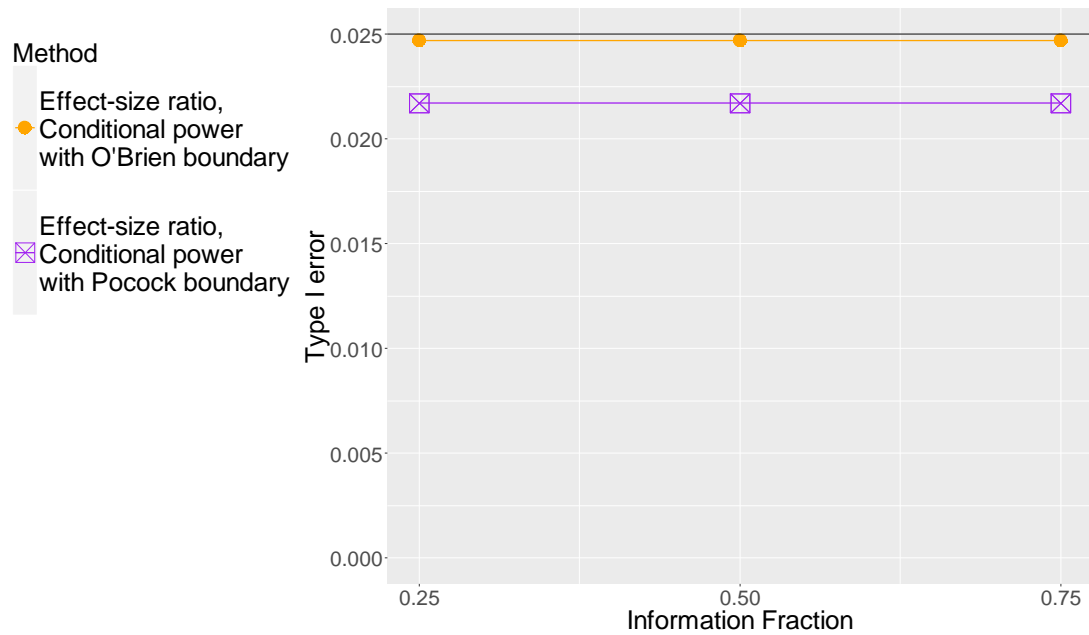


**Figure 7.15.** Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

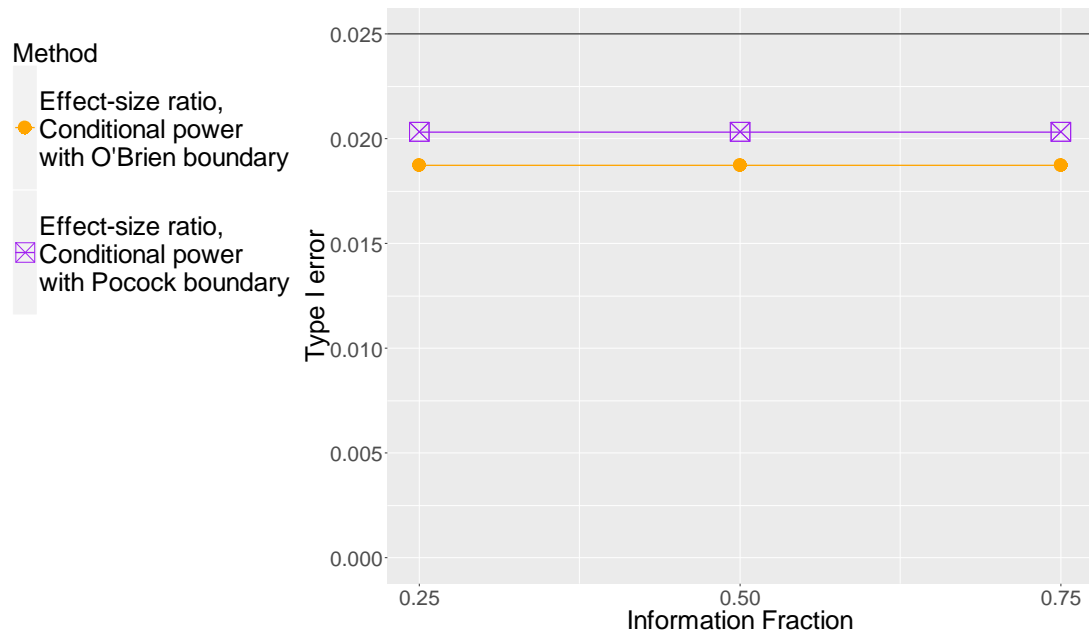


**Figure 7.16.** Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

### Second scenario of hazard ratio (i.e. 0.845)



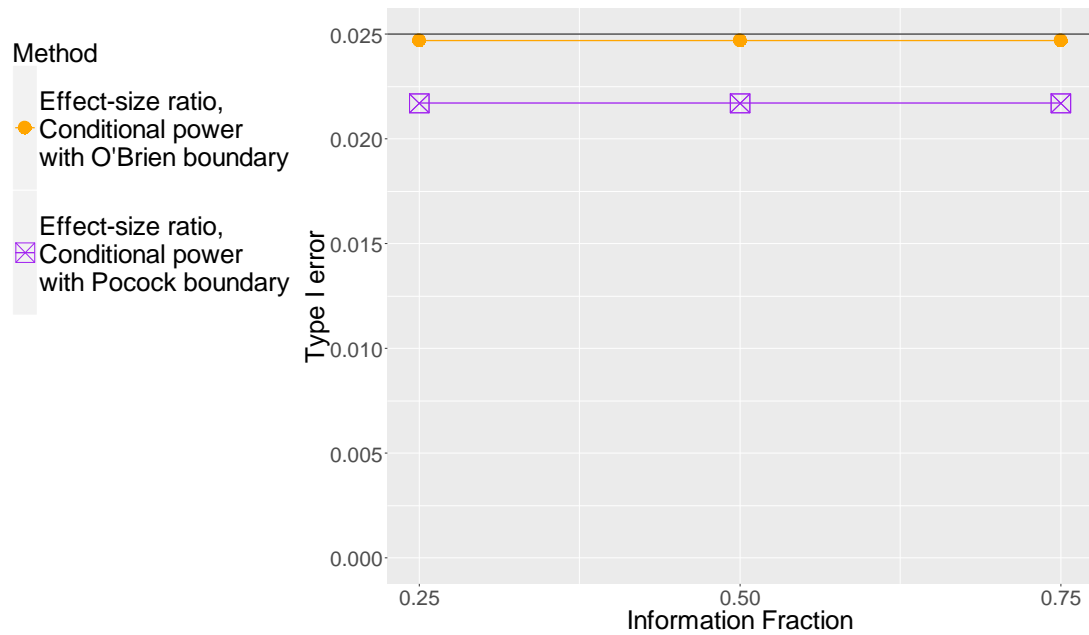
**Figure 7.17.** Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).



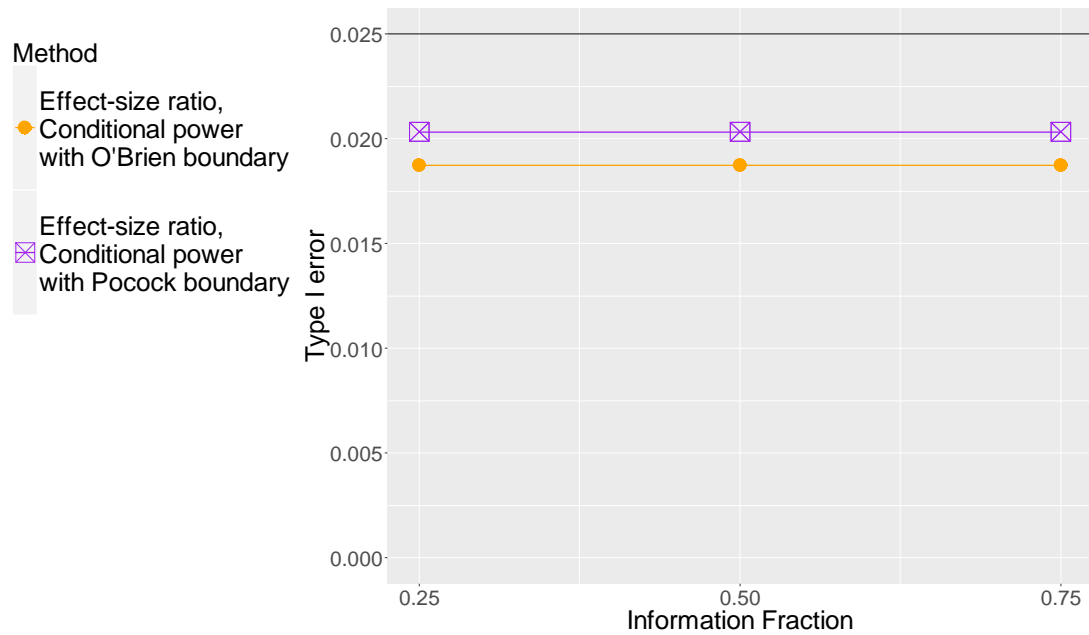
**Figure 7.18.** Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).



### Third scenario of hazard ratio (i.e. 0.807)

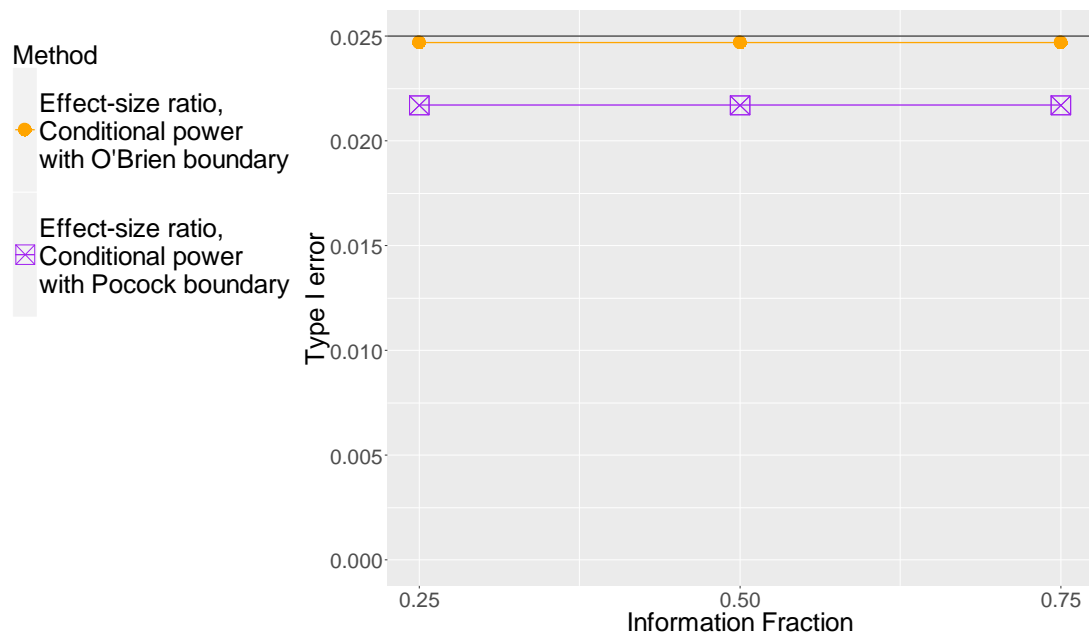


**Figure 7.19.** Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

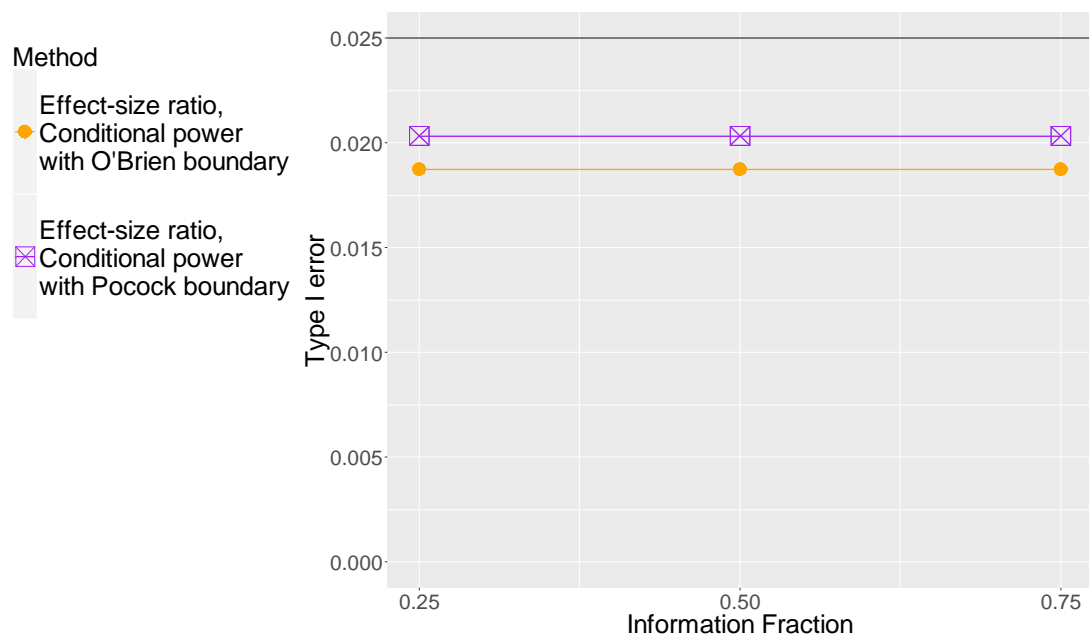


**Figure 7.20.** Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

#### Fourth scenario of hazard ratio (i.e. 0.765)



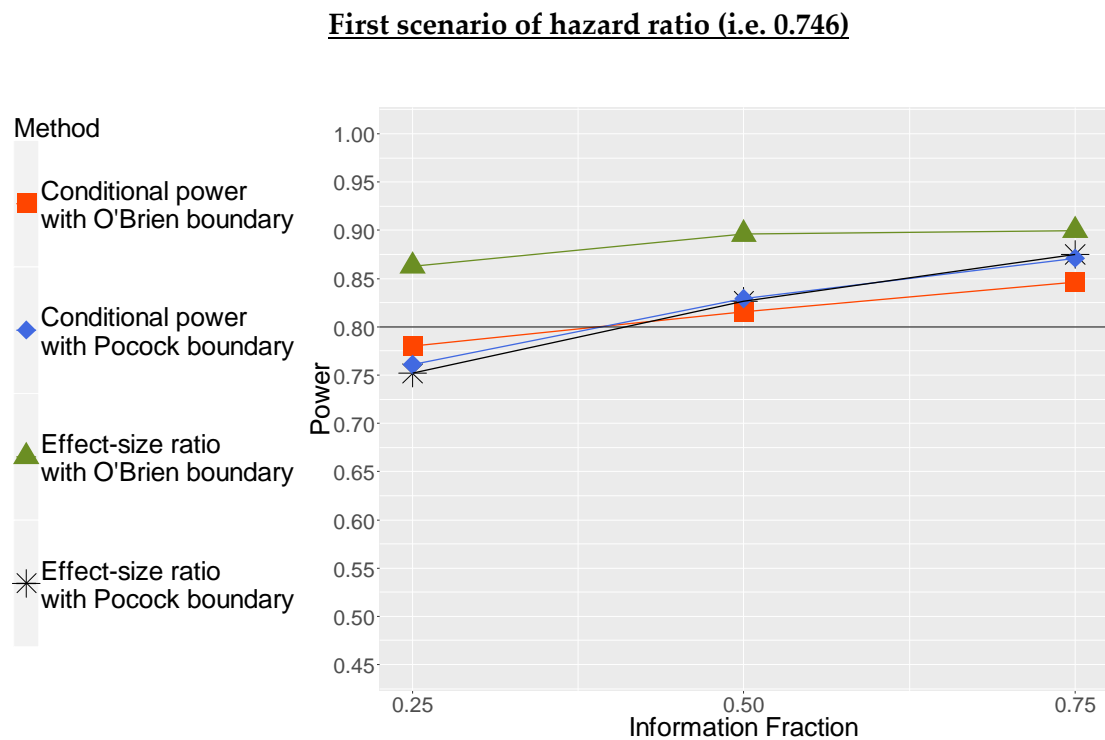
**Figure 7.21.** Type I error versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).



**Figure 7.22.** Type I error versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. The horizontal line represents the level of significance of the non-adaptive design (i.e. 0.025).

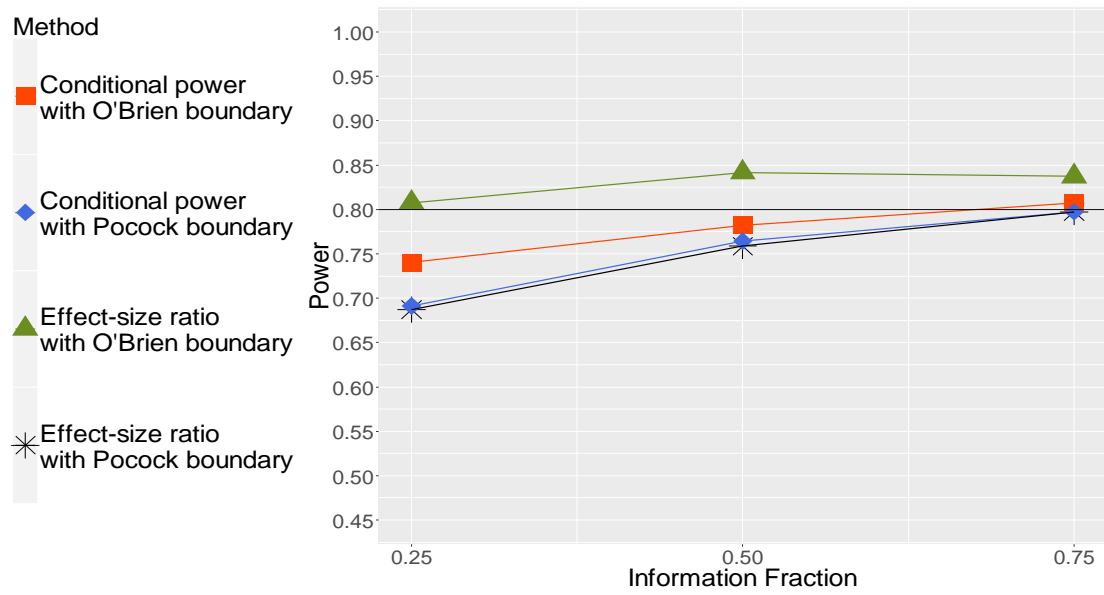
Figures 7.23 to 7.26 from Simulation study 1, under the alternative hypothesis, when the effect-size ratio method for sample size adjustment is applied, show that we can achieve greater power with the O'Brien-Fleming decision boundaries as compared to the Pocock's type boundaries in all scenarios of hazard ratio. Under conditional power method, in the first and fourth scenario of hazard ratio (i.e. 0.746 and 0.765 respectively) the power is higher with the O'Brien-Fleming decision boundaries compared to Pocock boundaries only at 25% information fraction (Figures 7.23, 7.26). For 50% and 75% information fraction, it seems that Pocock boundaries yield higher power as compared to the O'Brien-Fleming efficacy boundaries (Figures 7.23, 7.26). In the second scenario of hazard ratio (i.e. 0.845), when conditional power method is applied, the O'Brien-Fleming decision boundaries result in higher power compared to the Pocock boundaries in all cases of information fraction (Figure 7.24). In the third scenario of hazard ratio (i.e. 0.807), the conditional power method achieves higher power with the Pocock boundaries compared to the same method with the O'Brien-Fleming decision boundaries only at 75% information fraction (Figure 7.25). From Figure 7.23 which corresponds to the first scenario of hazard ratio (i.e. 0.746), we can see that at 50% and 75% information fraction, the power is improved compared to the nominal level of power in the non-adaptive design as it exceeds 80% in both types of sample size adjustment methods and both types of stopping boundaries. At 25% information fraction, the power exceeds the nominal level (i.e. 80%) only when the effect-size ratio method with the O'Brien-Fleming decision boundaries is applied. From Figure 7.24 which corresponds to the second scenario of hazard ratio (i.e. 0.843), it is shown that the power is improved compared to the nominal level of power in the non-adaptive design at 25% and 50% information fraction only when the effect-size ratio method with the O'Brien-Fleming boundaries is applied. At 75% information fraction, the power is improved compared to that of the fixed design only when the O'Brien-Fleming boundaries are applied. In the third scenario of hazard ratio (i.e. 0.807), the power exceeds the nominal level (i.e. 80%) in both sample size re-estimation methods and stopping boundaries at 75%

information fraction, whereas at 25% and 50% information fraction, only the effect-size ratio method with the O'Brien-Fleming boundaries achieves power greater than 80% (Figure 7.25). In the fourth scenario of hazard ratio (i.e. 0.765), we have greater power than 80% at 50% and 75% information fraction (Figure 7.26). At 25% information fraction, only the effect-size ratio method with the O'Brien-Fleming boundaries exceeds the 80% nominal level of power.



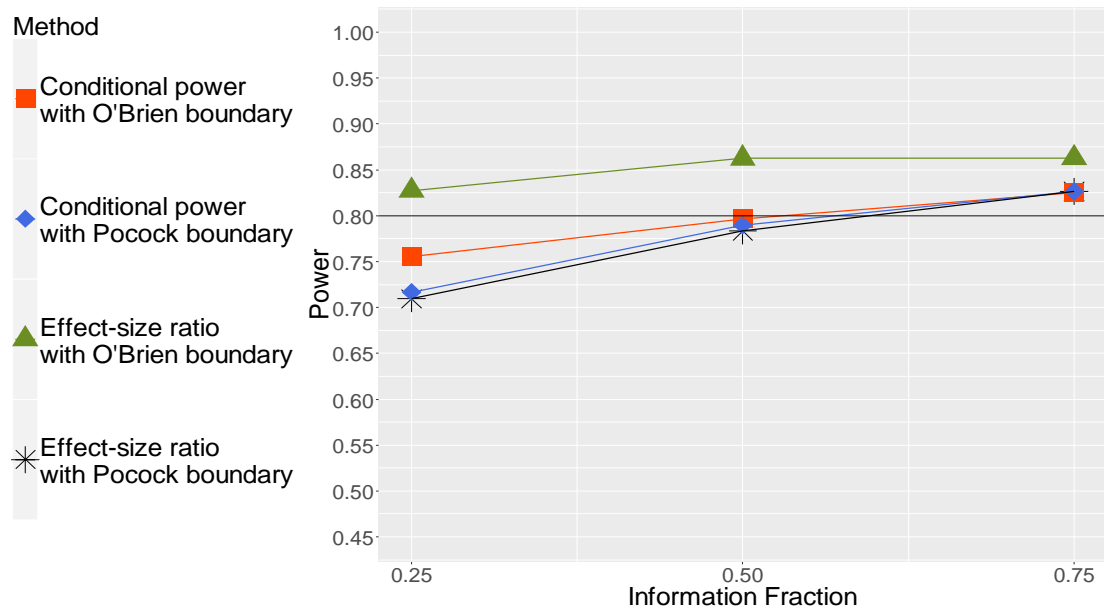
**Figure 7.23.** Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.

### Second scenario of hazard ratio (i.e. 0.845)



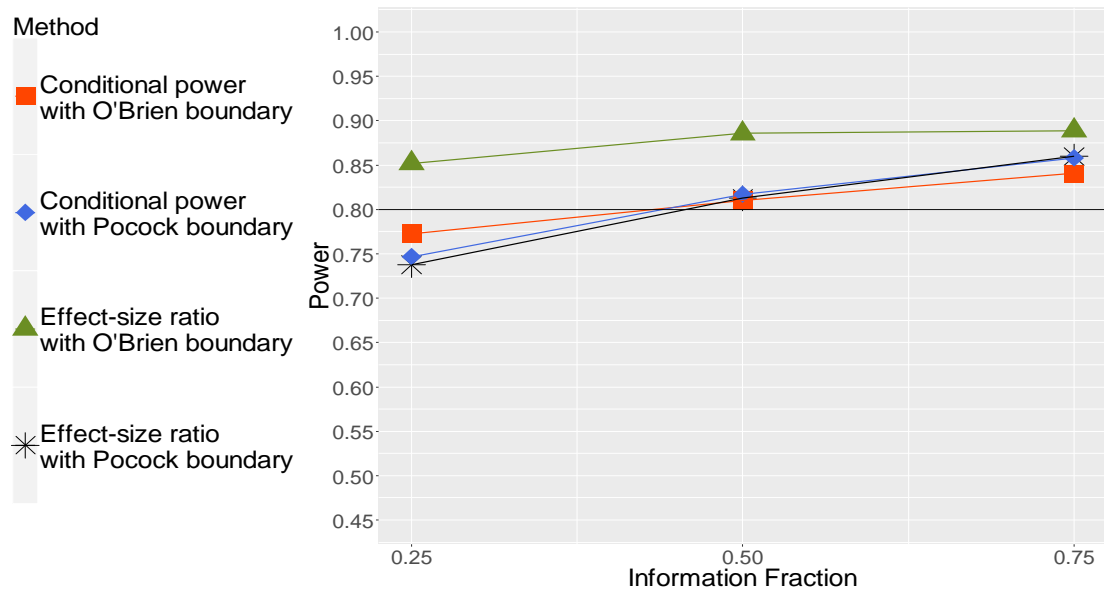
**Figure 7.24.** Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.

### Third scenario of hazard ratio (i.e. 0.807)



**Figure 7.25.** Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.

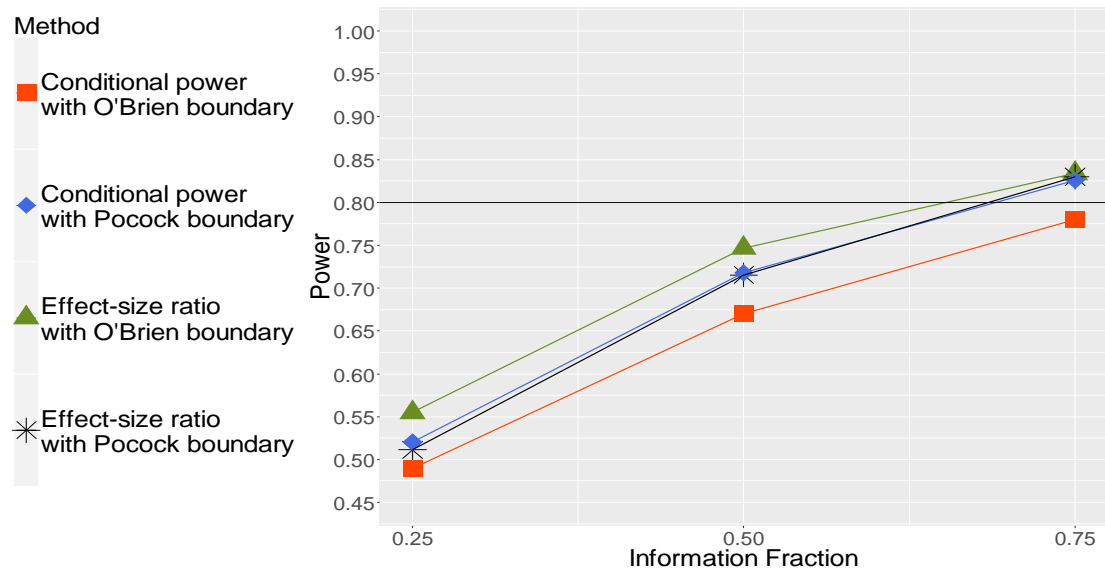
#### Fourth scenario of hazard ratio (i.e. 0.765)



**Figure 7.26.** Power versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the power of the non-adaptive design.

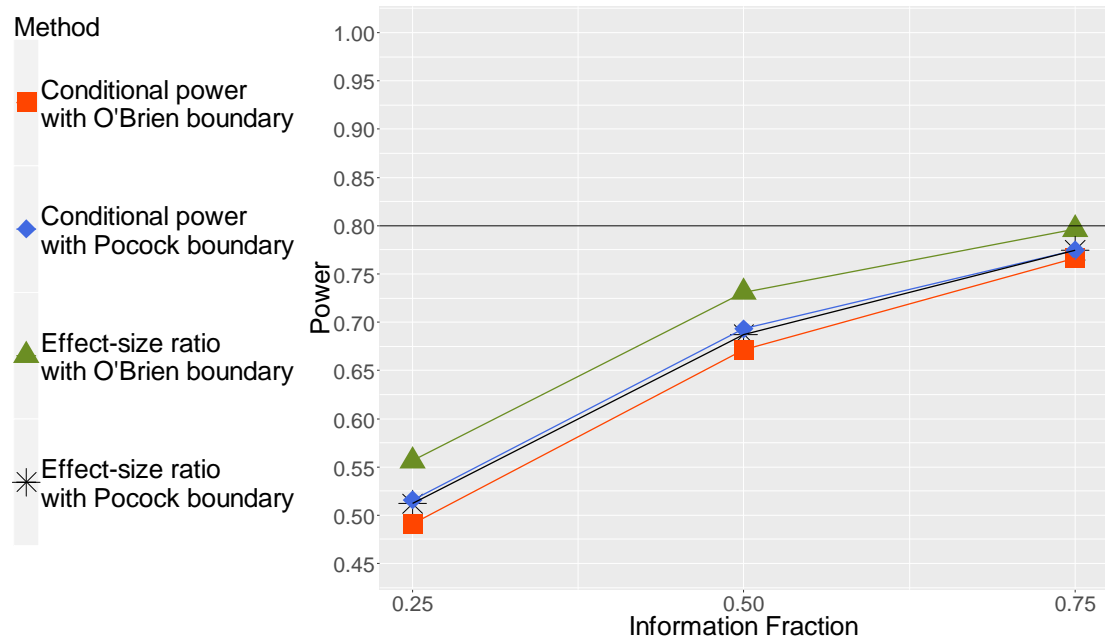
Results of power versus the information fraction under the alternative hypothesis of Simulation study 2 in each scenario of hazard ratio are presented in Figures 7.27 to 7.30 below. Under the alternative hypothesis, in the first, third and fourth scenario of hazard ratio (i.e. 0.746, 0.807 and 0.765 respectively), the power exceeds the nominal level (i.e. 80%) only at 75% information fraction when all types of sample size adjustment methods and stopping boundaries are used apart from the combination of the conditional power method with the O'Brien-Fleming decision boundaries (Figures 7.27, 7.29, 7.30). Among the different sample size adjustment methods and stopping boundaries, the effect-size ratio method with the O'Brien-Fleming decision boundaries results in the highest power and the conditional power method with the same type of stopping boundaries results in the smallest power (Figure 7.27, 7.28) in all cases of hazard ratio. In the second scenario of hazard ratio (i.e. 0.845), all resulting power are below the horizontal line which corresponds to the nominal level of power (i.e. 80%), thus, the power is not preserved (Figure 7.28).

### First scenario of hazard ratio (i.e. 0.746)



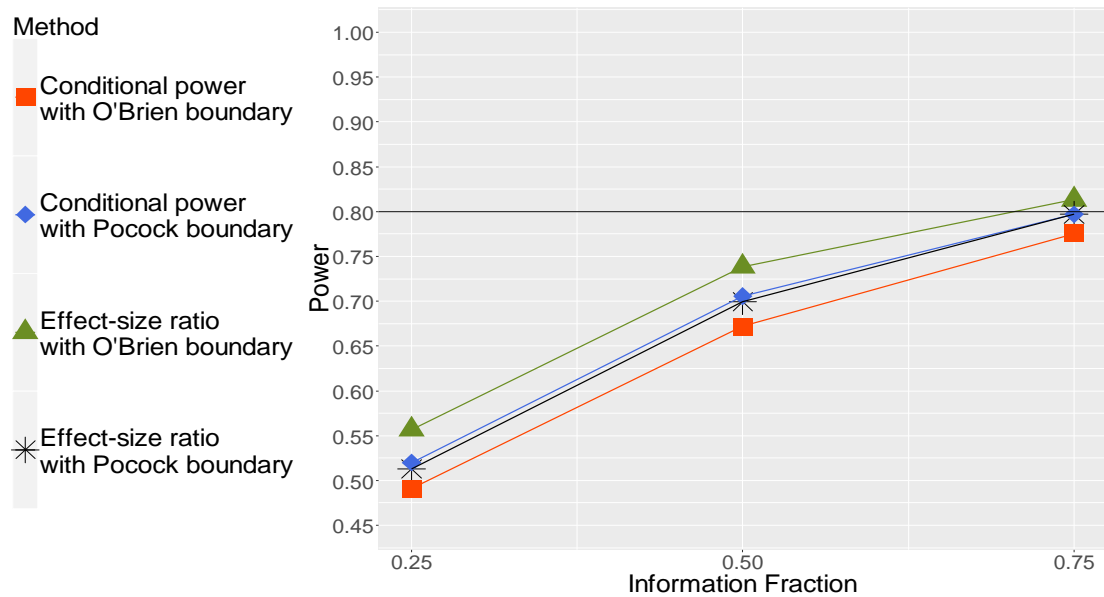
**Figure 7.27.** Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio (i.e. 0.746). The horizontal line represents the power of the non-adaptive design.

### Second scenario of hazard ratio (i.e. 0.845)



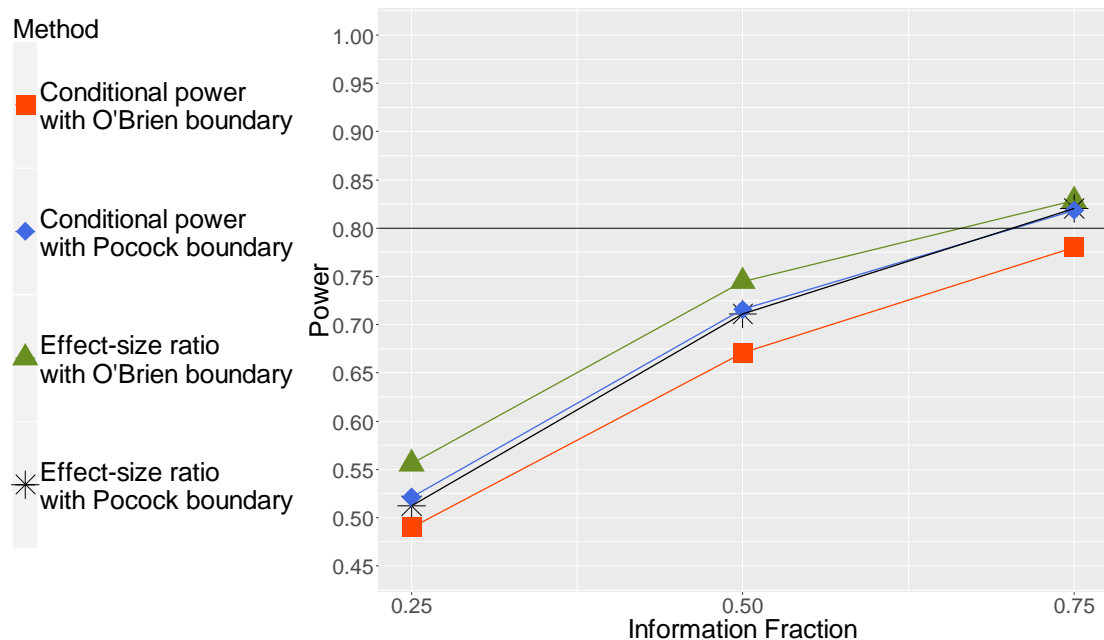
**Figure 7.28.** Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio (i.e. 0.807). The horizontal line represents the power of the non-adaptive design.

### Third scenario of hazard ratio (i.e. 0.807)



**Figure 7.29.** Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio (i.e. 0.845). The horizontal line represents the power of the non-adaptive design.

### Fourth scenario of hazard ratio (i.e. 0.765)

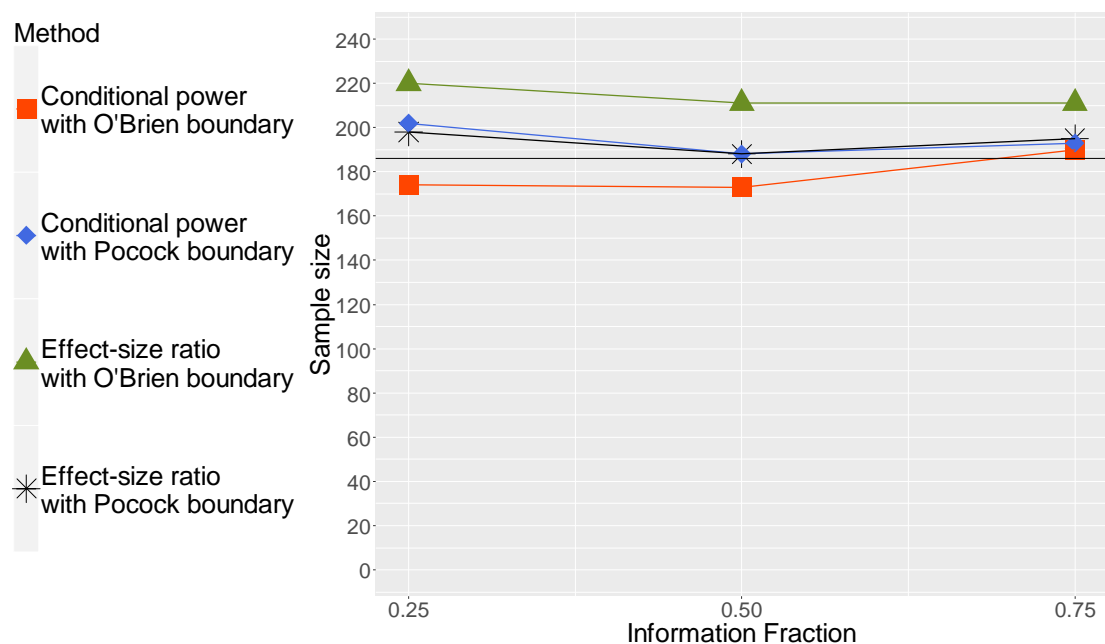


**Figure 7.30.** Power versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the power of the non-adaptive design.



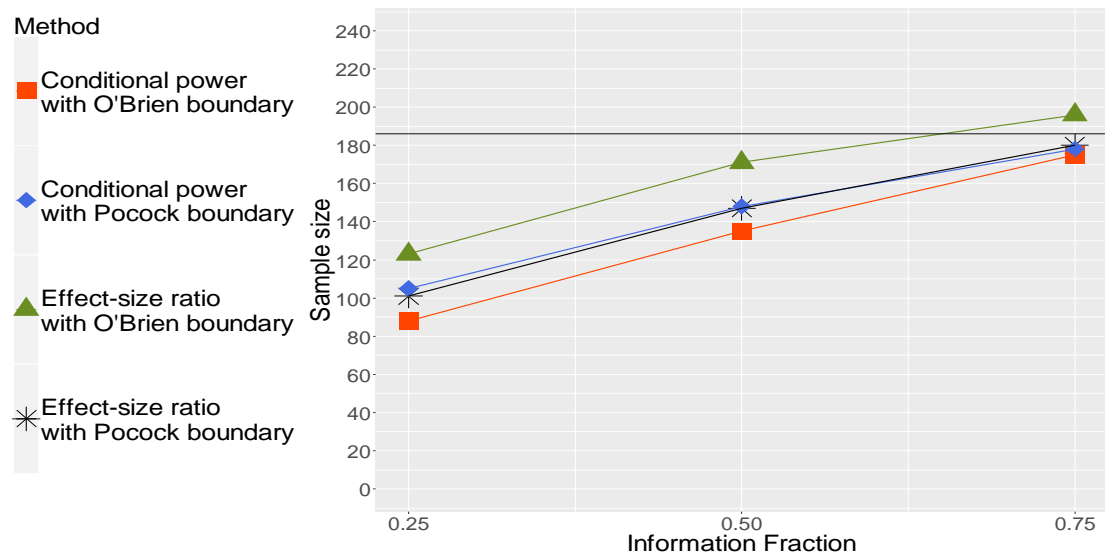
### 7.5.2.3. Sample size of the study

In Simulation study 1, under the alternative hypothesis of the first scenario of hazard ratio (i.e. 0.746), the horizontal line which indicates the sample size of the non-adaptive design in Figure 7.31, shows loss of efficiency at 75% information fraction with all the combinations of sample size recalculation methods and stopping boundaries.



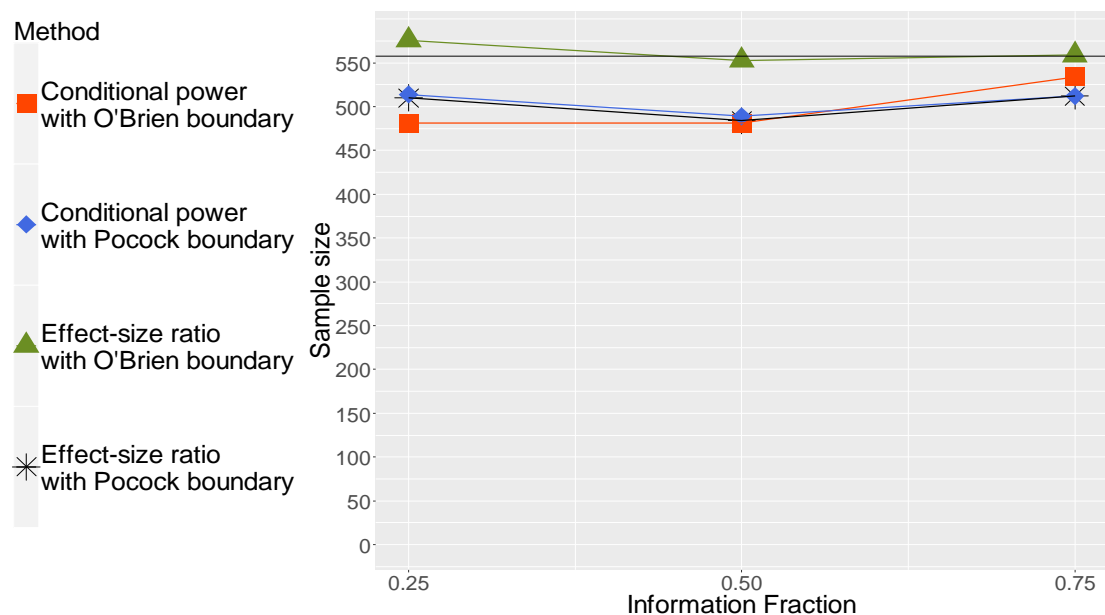
**Figure 7.31.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio (i.e. 0.746).

At 25% and 50% information fraction, there is gain in efficiency only in the case of the conditional power method with the O'Brien-Fleming boundaries as the corresponding values are below the horizontal line which indicates the sample size of the non-adaptive design (Figure 7.31). However, with the introduction of both efficacy and futility stopping in the trial (Simulation study 2), there is gain in efficiency in all cases apart from the effect-size ratio method with the O'Brien-Fleming boundaries at 75% information fraction as the values in Figure 7.32 are below the horizontal line.



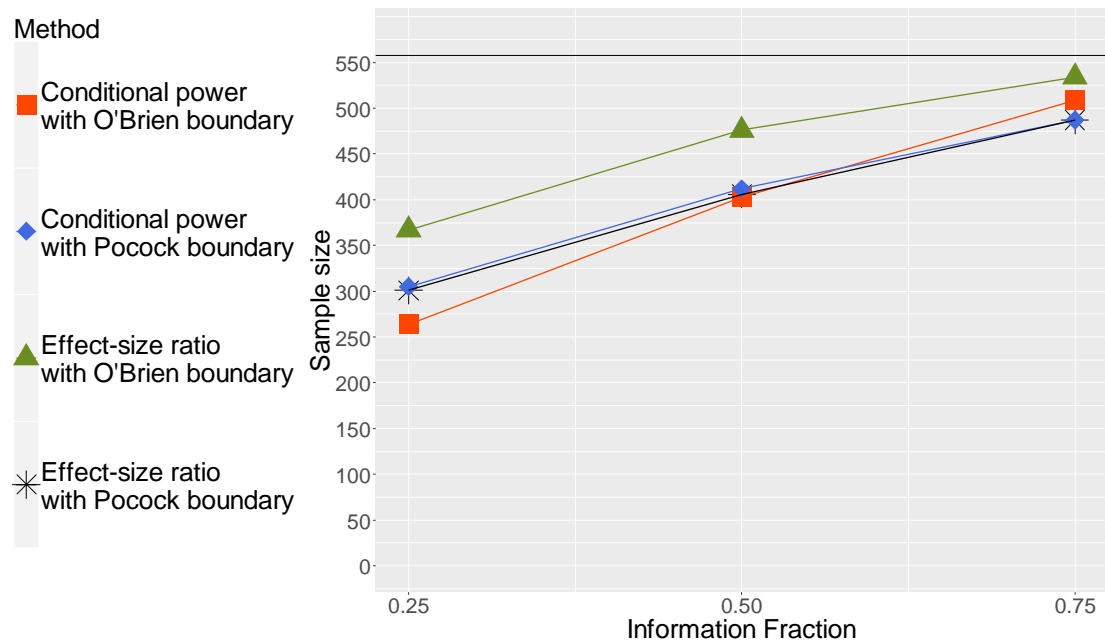
**Figure 7.32.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio (i.e. 0.746).

Under the alternative hypothesis of the second scenario of hazard ratio (i.e. 0.845), it can be seen in Figure 7.33 that there is gain in efficiency in all cases apart from the combination of effect-size ratio method with the O'Brien-Fleming boundaries at 25% and 75% of information fraction.



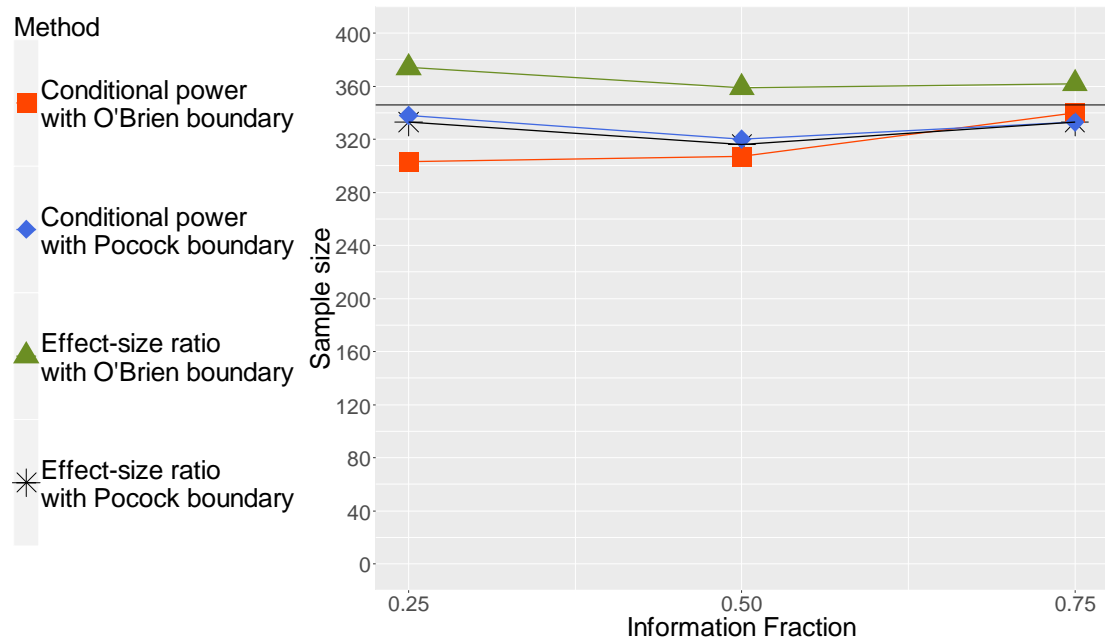
**Figure 7.33.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio (i.e. 0.845). The horizontal line represents the sample size of the non-adaptive design.

However, with the introduction of both efficacy and futility stopping of the trial (Simulation study 2), the sample size is decreased with all the combinations of sample size recalculation methods and stopping boundaries across the different percentages of information fraction compared to the sample size of the non-adaptive design (Figure 7.34).

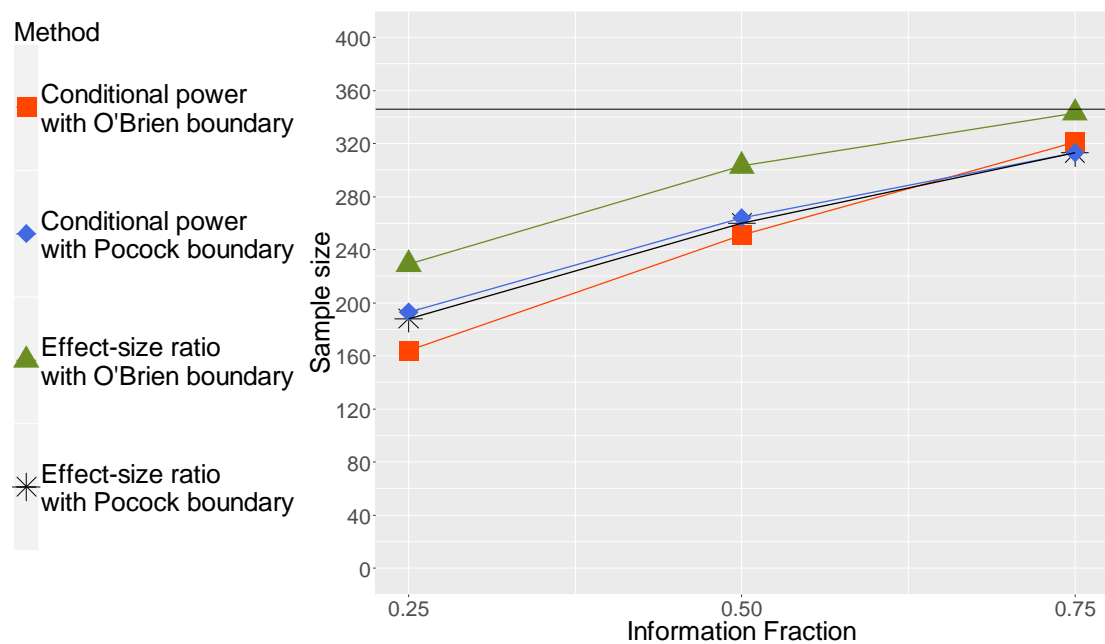


**Figure 7.34.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio (i.e. 0.845). The horizontal line represents the sample size of the non-adaptive design.

Under the alternative hypothesis of the third scenario of hazard ratio (i.e. 0.807), loss of efficiency can be observed at all levels of information fraction with the combination of the effect-size ratio method and the O'Brien-Fleming boundaries (Figure 7.35), whereas in Simulation study 2, all values in Figure 7.36 are below the horizontal line and thus only gain in efficiency can be observed.



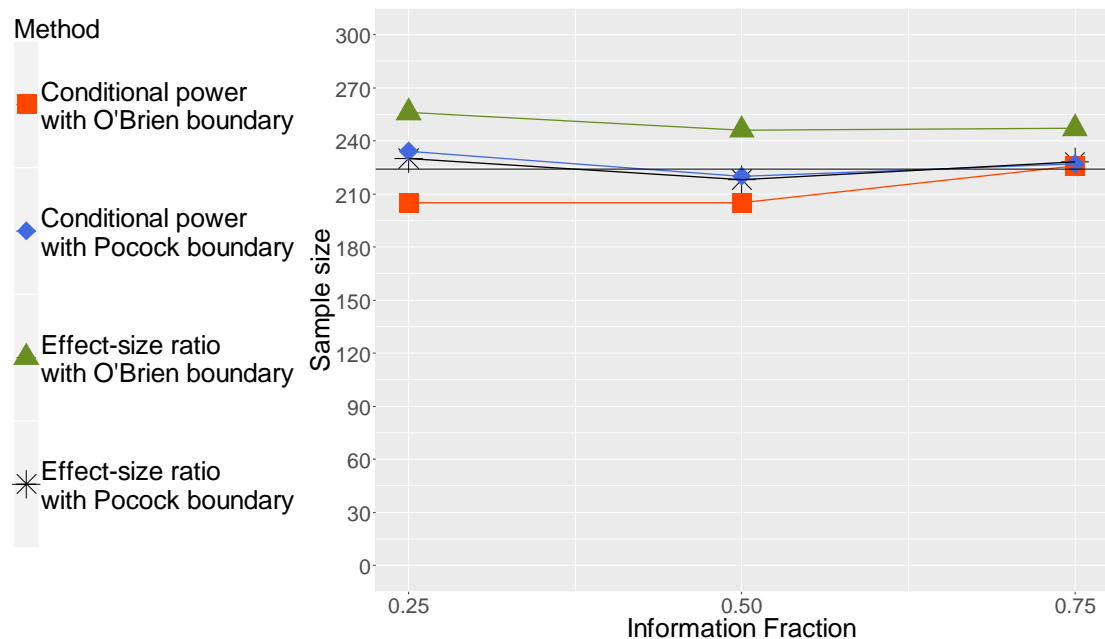
**Figure 7.35.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio (i.e. 0.807). The horizontal line represents the sample size of the non-adaptive design.



**Figure 7.36.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio (i.e. 0.807). The horizontal line represents the sample size of the non-adaptive design.

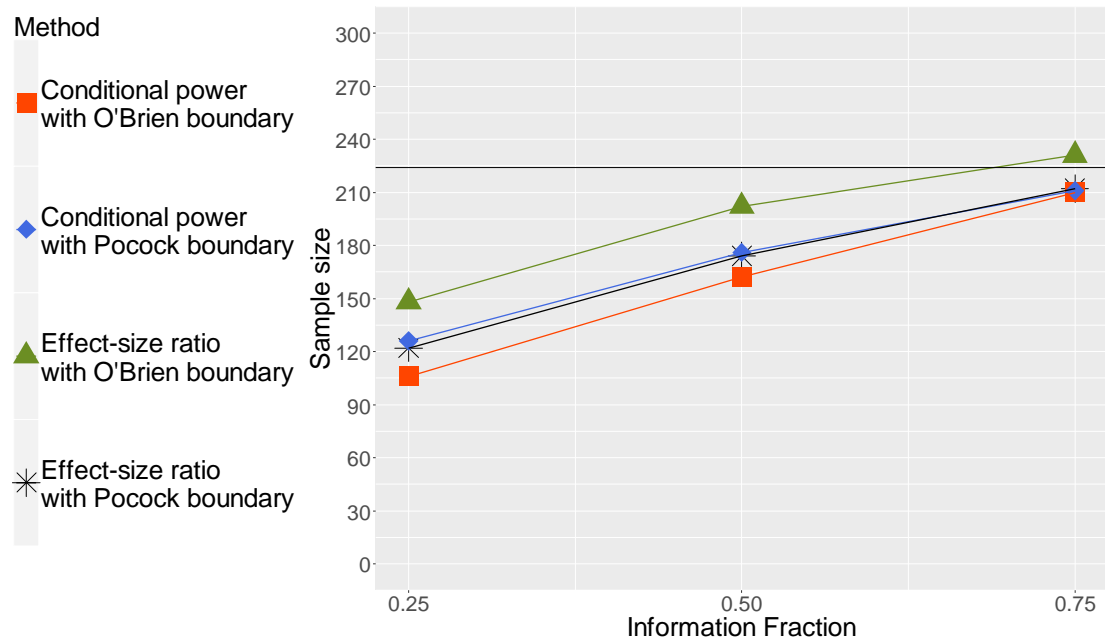
Under the alternative hypothesis of the fourth scenario of hazard ratio (i.e. 0.765), the sample size is increased compared to the sample size of the non-adaptive design at 75% information fraction (Figure 7.37). However, at 50% information

fraction, the only loss of efficiency is observed with the effect-size ratio method and the O'Brien-Fleming boundaries and at 25% information fraction, the only gain in efficiency is obtained with the conditional power method and the O'Brien-Fleming boundaries (Figure 7.37).



**Figure 7.37.** Sample size versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the sample size of the non-adaptive design.

In Simulation study 2 where we allow the trial to stop either for efficacy or futility, the sample size is smaller than that of the non-adaptive design apart from the case of the effect-size ratio method with the O'Brien-Fleming boundaries at 75% information fraction (Figure 7.38).

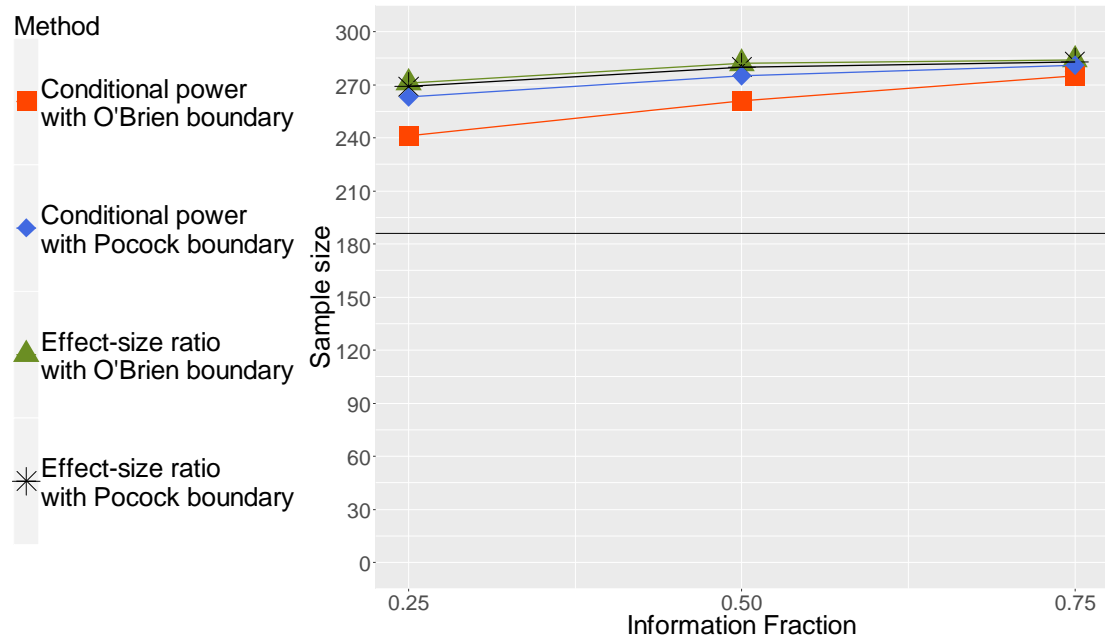


**Figure 7.38.** Sample size versus the information fraction under the alternative and null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio (i.e. 0.765). The horizontal line represents the sample size of the non-adaptive design.

In Simulation study 1, under the null hypothesis of all scenarios of hazard ratio, all the values are above the horizontal line in Figures 7.39 to 7.42 indicating the increase of the sample size compared to the sample size of the non-adaptive design. In both methods of sample size recalculation and both types of decision boundaries, the number of patients increases when the information fraction also increases in each case of hazard ratio. In each scenario of hazard ratio, the largest increase of the sample size as compared to the number of patients required for the non-adaptive approach corresponds to the effect-size ratio method with O'Brien-Fleming decision boundaries and 75% information fraction. More precisely, in the first scenario of hazard ratio (i.e. 0.746) from 186 patients per arm which are required for the non-adaptive approach, we have calculated 284 patients per arm which is very close to the upper limit of sample size for the study which we allowed (i.e. 286 patients per arm). For the second scenario of hazard ratio (i.e. 0.845), from 558 patients per arm required for the non-adaptive approach, we have calculated 559 patients per arm (the upper limit of 658 patients per arm was allowed). For the third scenario of hazard ratio, (i.e. 0.807), from 346 patients per arm required for the fixed design, we have calculated 362 patients per arm allowing a maximum sample size of 446 patients per

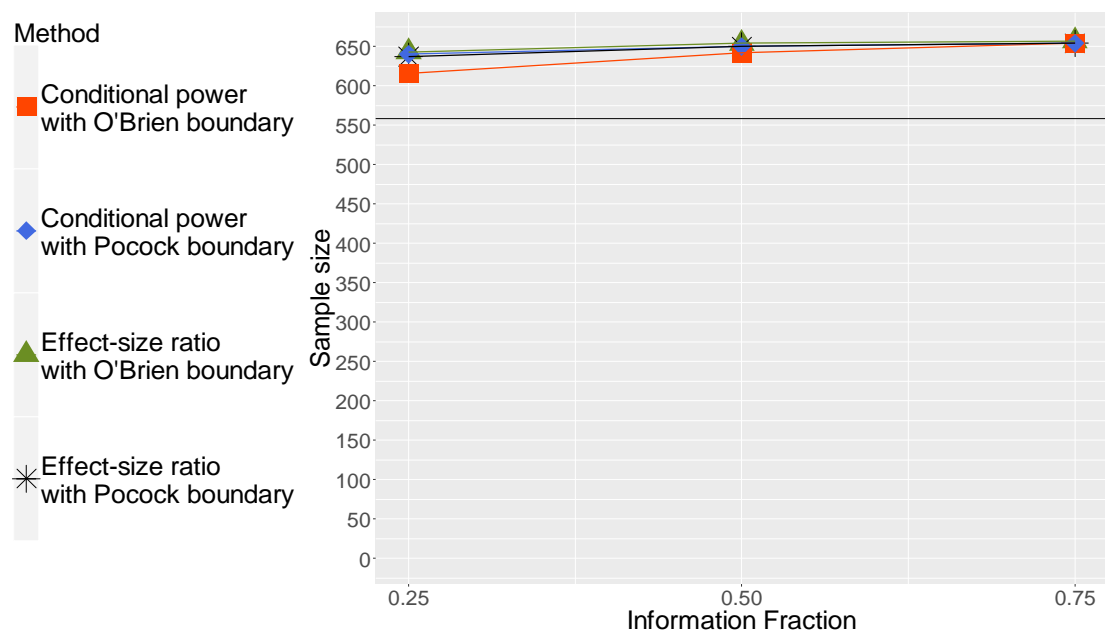
arm. For the fourth scenario of hazard ratio, (i.e. 0.765), from 224 patients per arm required for the fixed design, we reached 247 patients per arm in the adaptive approach allowing for a maximum sample size of 324 patients.

#### First scenario of hazard ratio (i.e. 0.746)



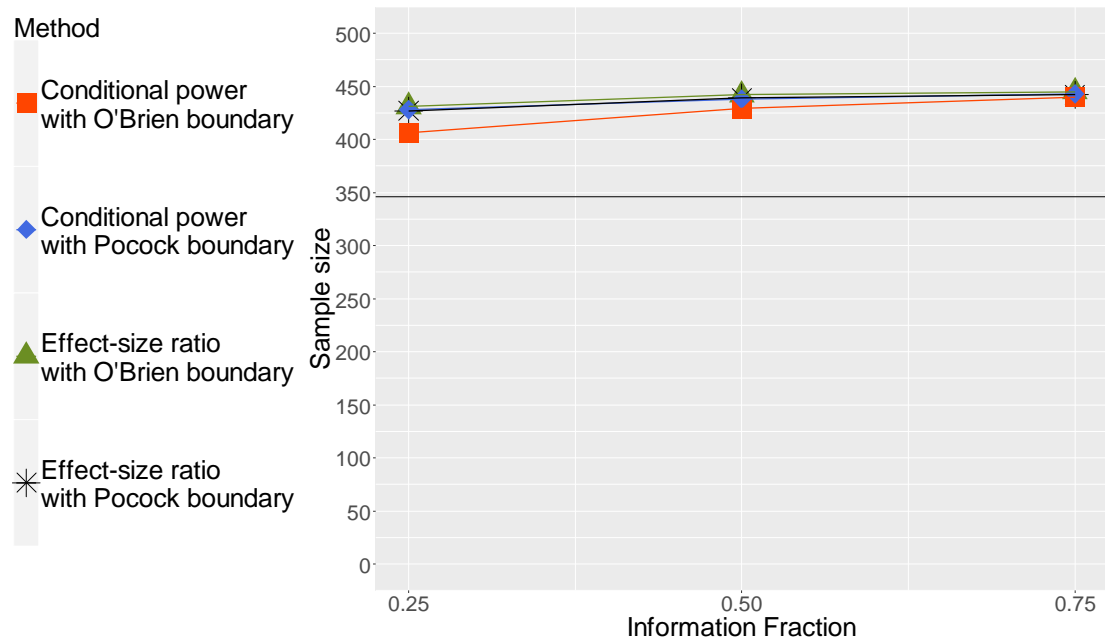
**Figure 7.39.** Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

#### Second scenario of hazard ratio (i.e. 0.845)



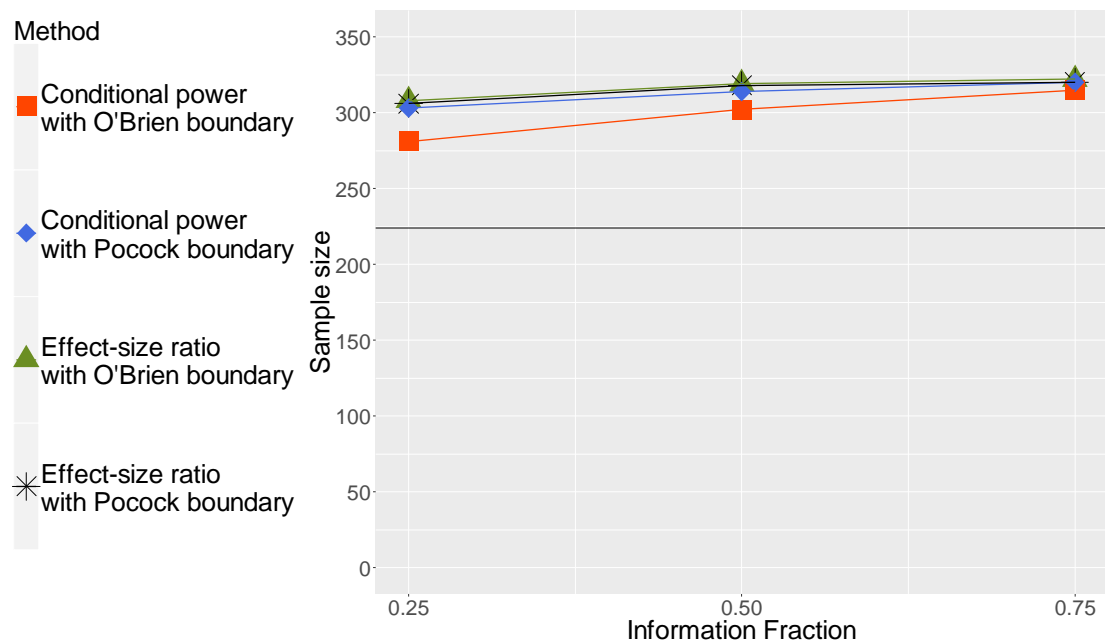
**Figure 7.40.** Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

**Third scenario of hazard ratio (i.e. 0.807)**



**Figure 7.41.** Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

**Fourth scenario of hazard ratio (i.e. 0.765)**

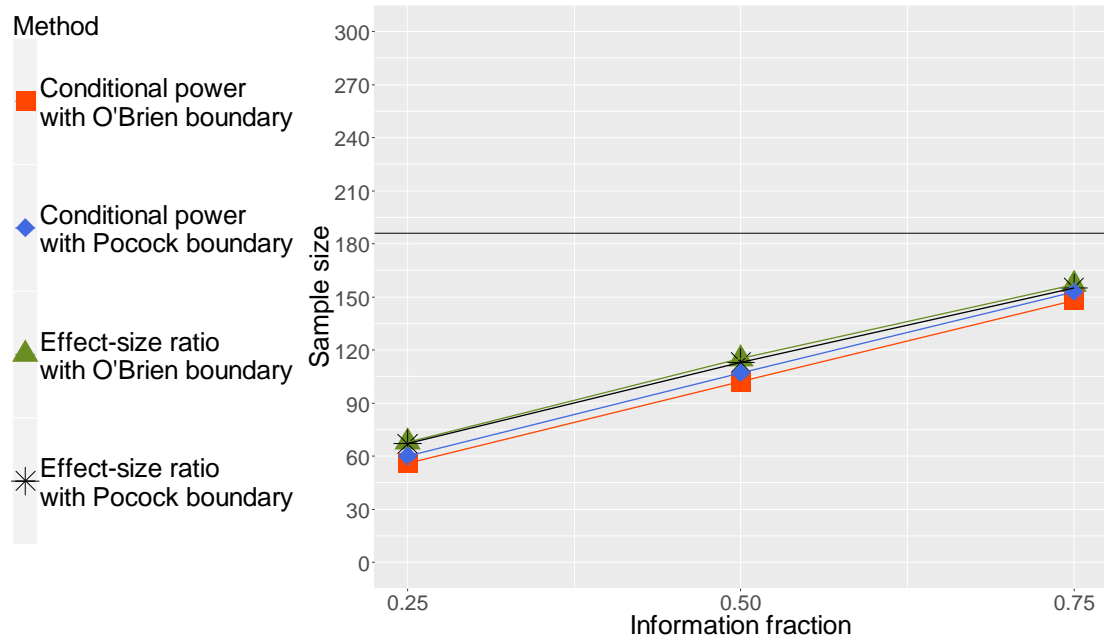




**Figure 7.42.** Sample size versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

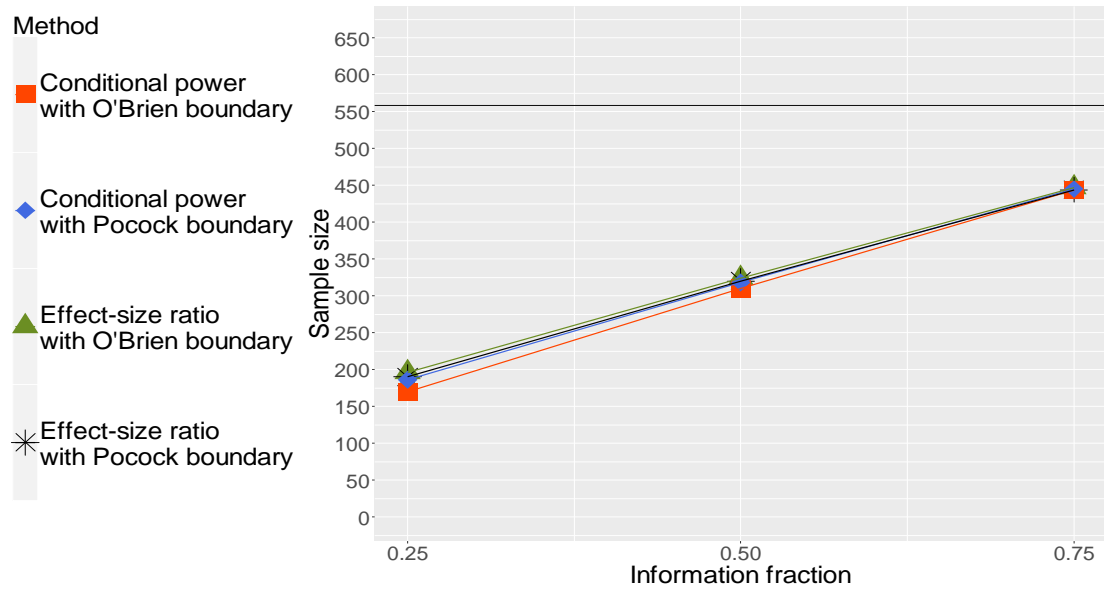
In Simulation study 2, under the null hypothesis of all scenarios of hazard ratio, all the values are below the horizontal line in Figures 7.43 to 7.46 indicating the decrease of the sample size compared to the sample size of the non-adaptive design. Hence, we can observe gain in efficiency when allowing the trial to stop early either for efficacy or futility. In both methods of sample size recalculation and both types of decision boundaries for all scenarios of hazard ratios, the number of patients increases when the information fraction increases. In each scenario of hazard ratio, the largest increase of the sample size as compared to the number of patients required for the non-adaptive approach corresponds to the effect-size ratio method with O'Brien-Fleming decision boundaries and 75% information fraction. More specifically, in the first scenario of hazard ratio, (i.e. 0.746), from 186 patients per arm which are required for the non-adaptive approach, we have calculated 196 patients per arm allowing for the upper limit of sample size at 286 patients per arm. For the second scenario of hazard ratio (i.e. 0.845), from 558 patients per arm which are required for the non-adaptive approach, we have calculated 534 patients per arm (the upper limit of 658 patients per arm was allowed). For the third scenario of hazard ratio (i.e. 0.807), from 346 patients per arm required for the fixed design, we have calculated 343 patients per arm allowing a maximum sample size of 446 patients per arm. For the last scenario of hazard ratio (i.e. 0.765), from 224 patients per arm required for the fixed design, we reached 231 patients per arm in the adaptive approach allowing for a maximum sample size of 324 patients.

### First scenario of hazard ratio (i.e. 0.746)



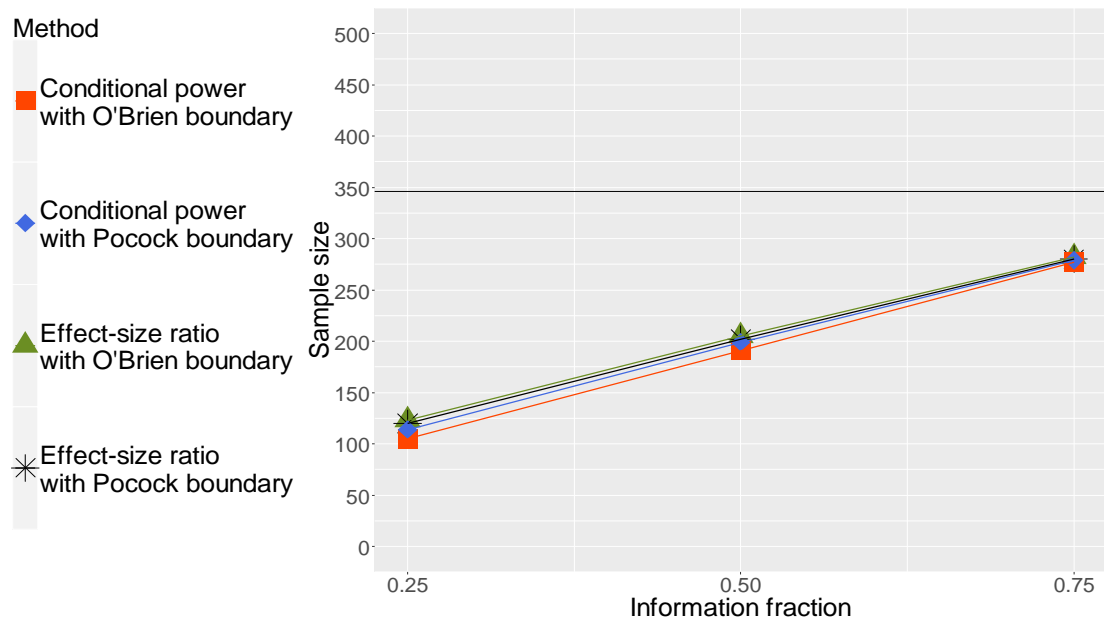
**Figure 7.43.** Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

### Second scenario of hazard ratio (i.e. 0.845)



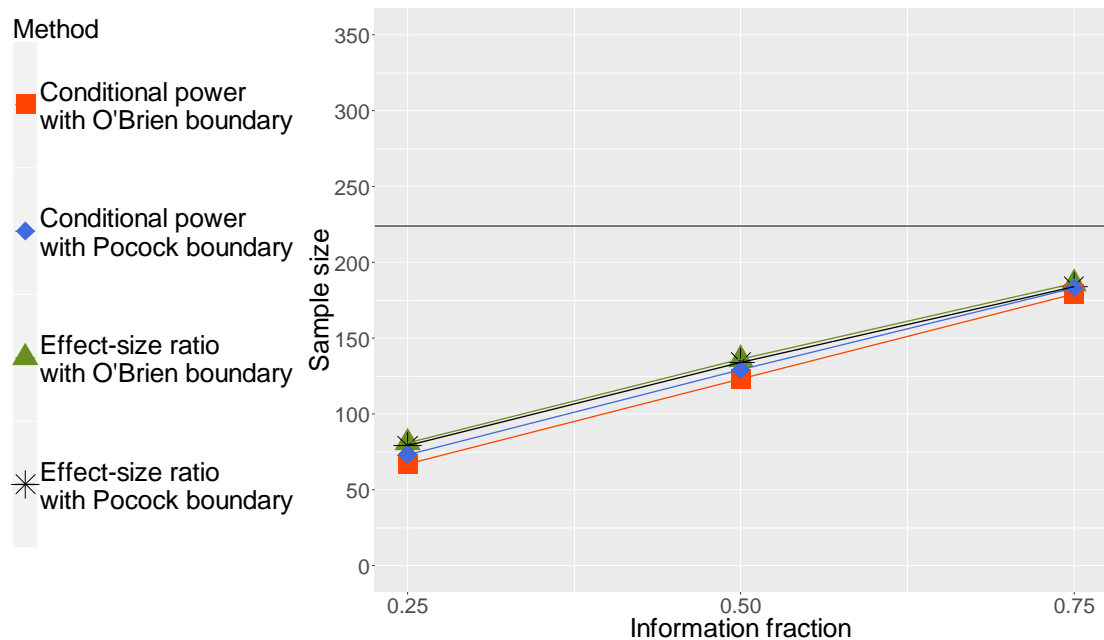
**Figure 7.44.** Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

### Third scenario of hazard ratio (i.e. 0.807)



**Figure 7.45.** Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

### Fourth scenario of hazard ratio (i.e. 0.765)

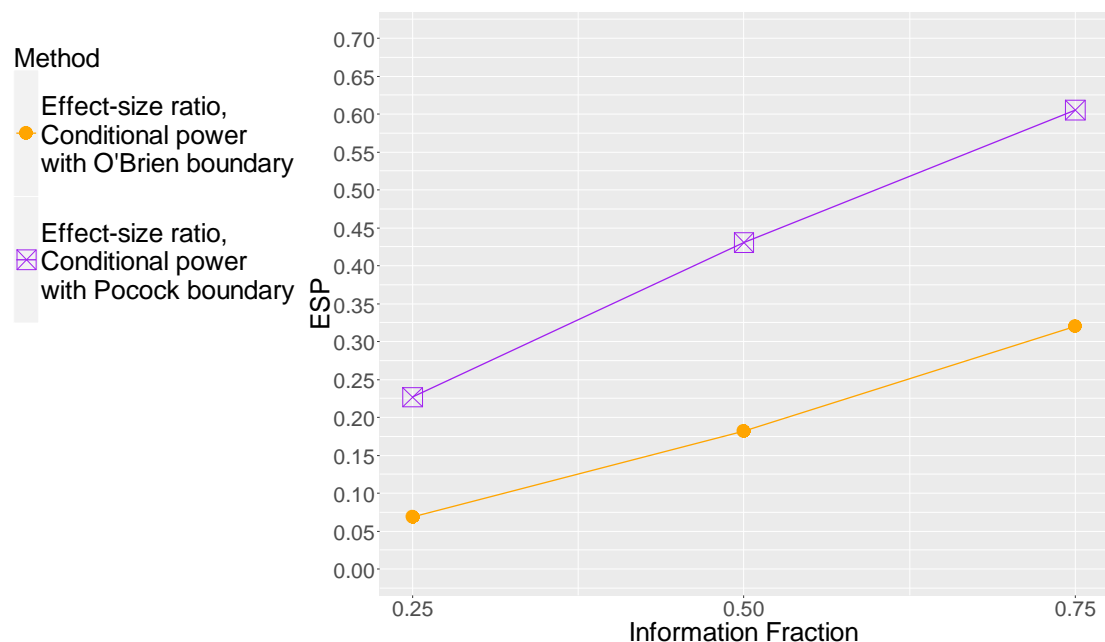


**Figure 7.46.** Sample size versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio. The horizontal line represents the sample size of the non-adaptive design.

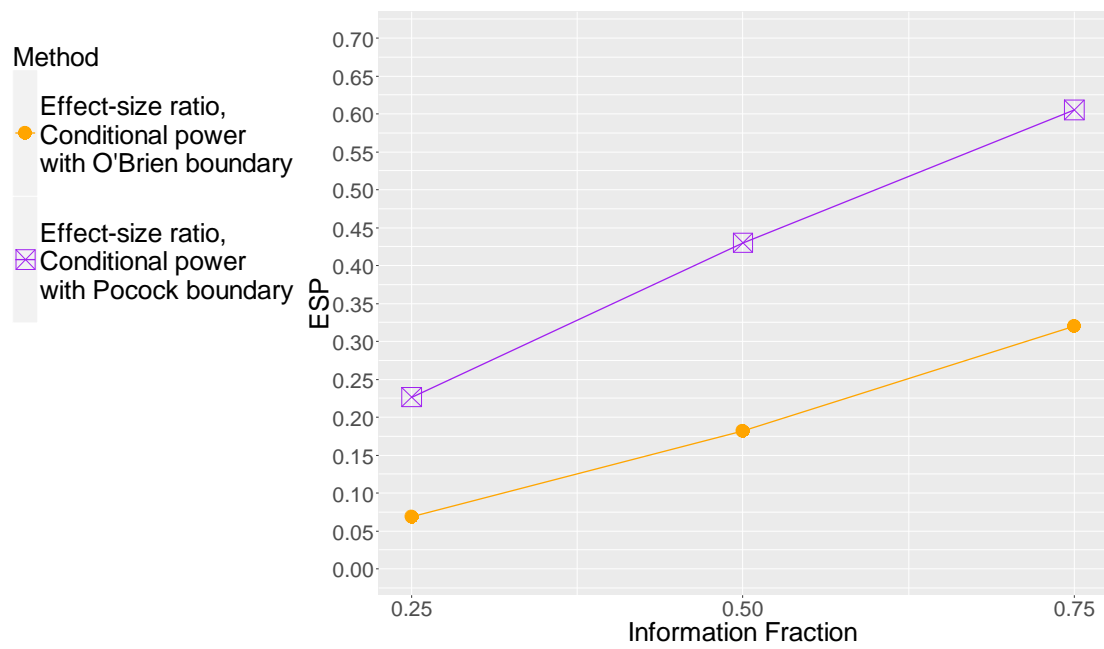
#### 7.5.2.4. Efficacy stopping probability

In both simulation studies, under the alternative hypothesis of all scenarios of hazard ratio, the efficacy stopping probability depends only on the type of stopping boundaries. It can be seen in Figures 7.47 to 7.54 that the efficacy stopping probability increases with the increase of information fraction. Consequently, the largest value is obtained at 75% information fraction. Additionally, the efficacy stopping probability is always greater with the Pocock decision boundaries compared to the probabilities obtained with the O'Brien-Fleming decision boundaries.

##### **First scenario of hazard ratio (i.e. 0.746)**

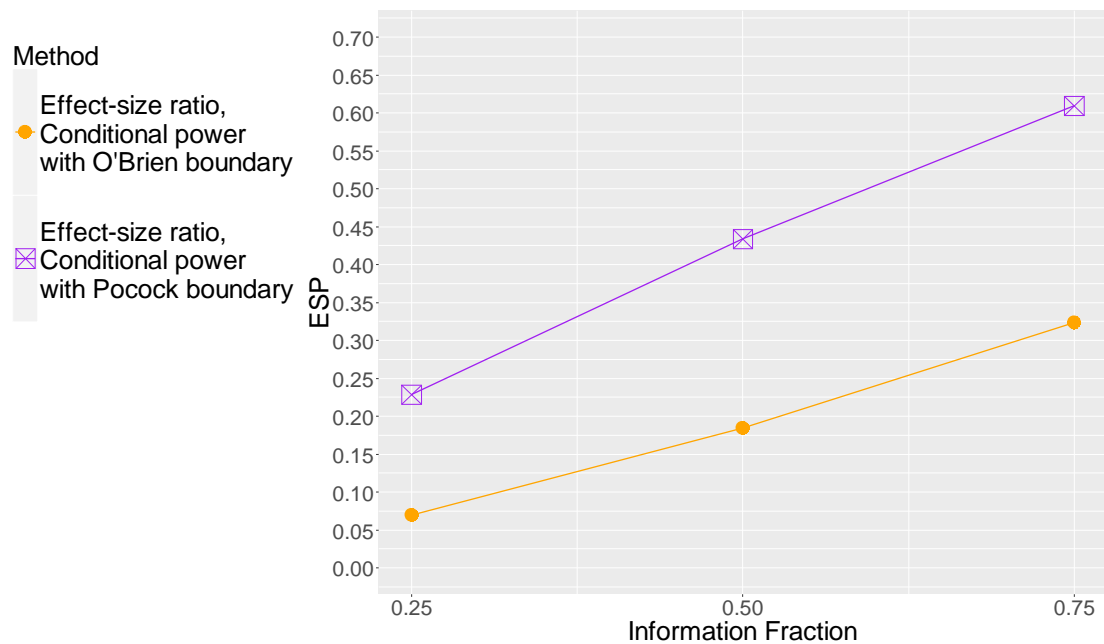


**Figure 7.47.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the first scenario of hazard ratio.

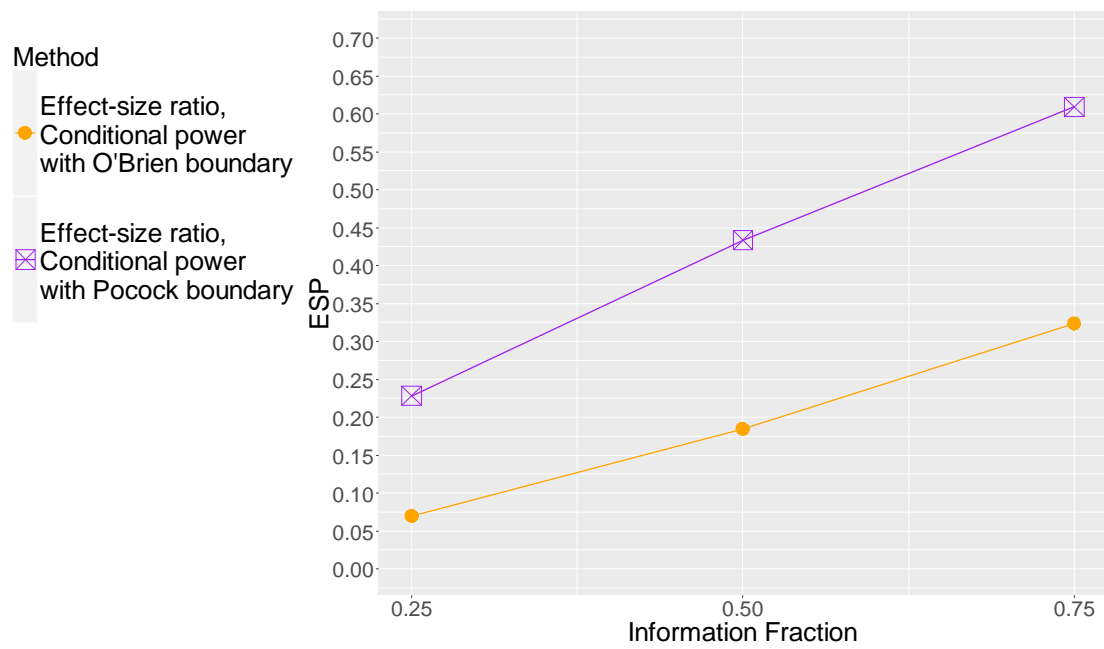


**Figure 7.48.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio.

### Second scenario of hazard ratio (i.e. 0.845)

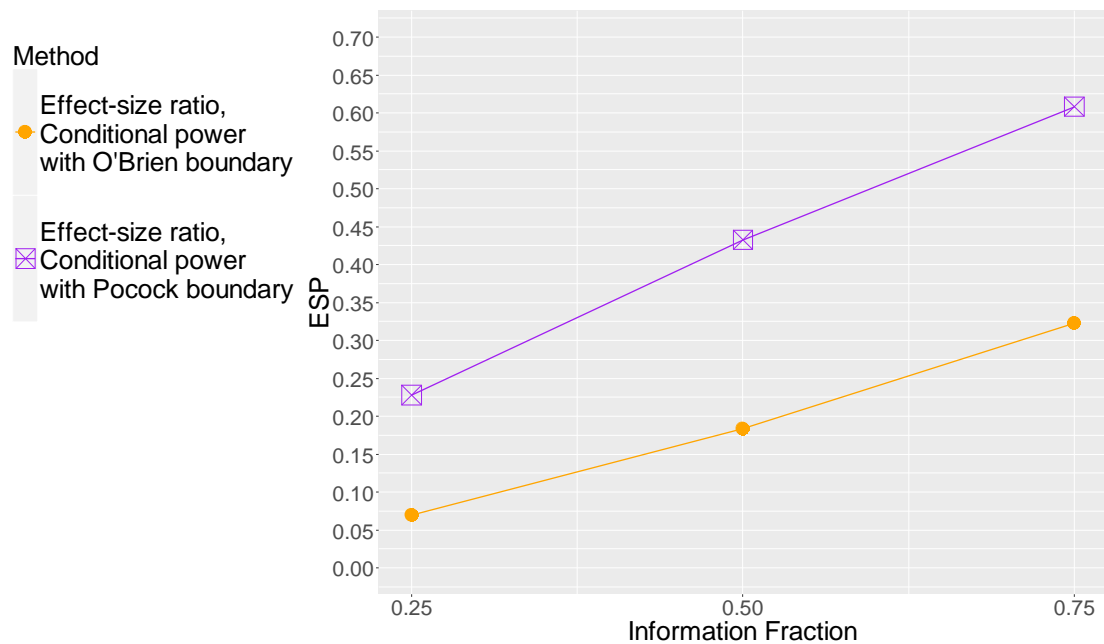


**Figure 7.49.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the second scenario of hazard ratio.

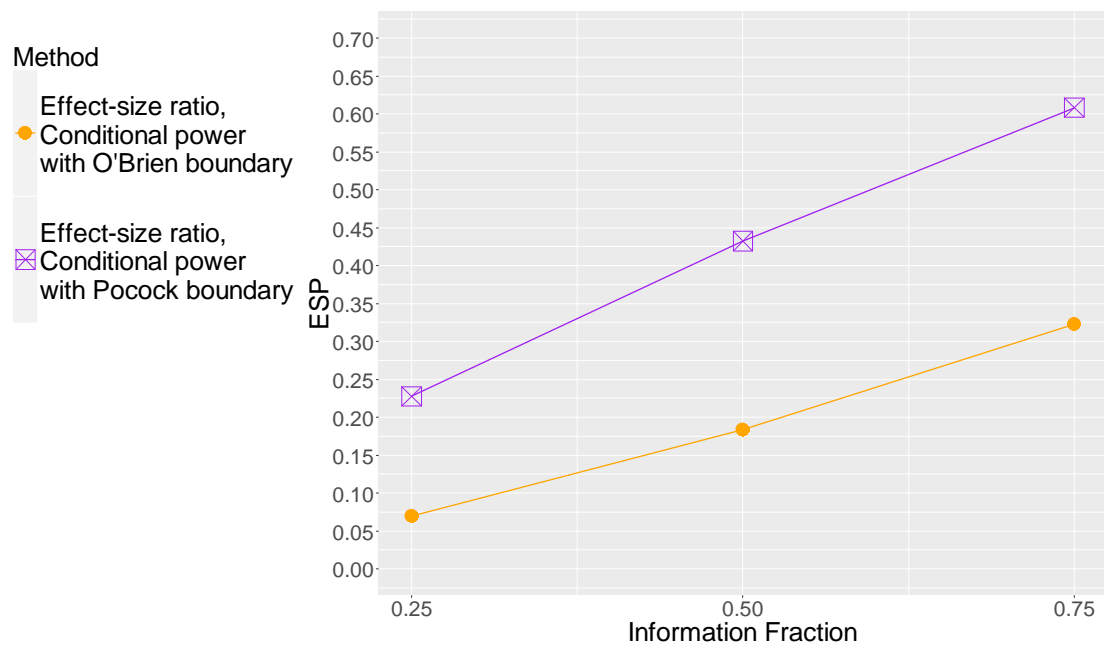


**Figure 7.50.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio.

### Third scenario of hazard ratio (i.e. 0.807)

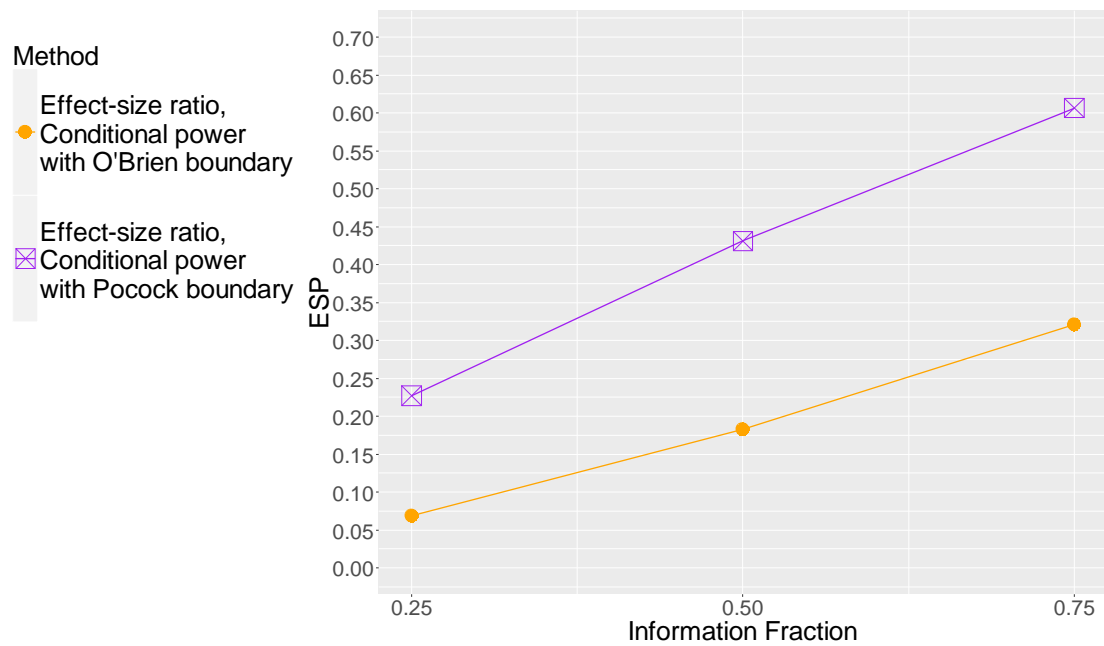


**Figure 7.51.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the third scenario of hazard ratio.

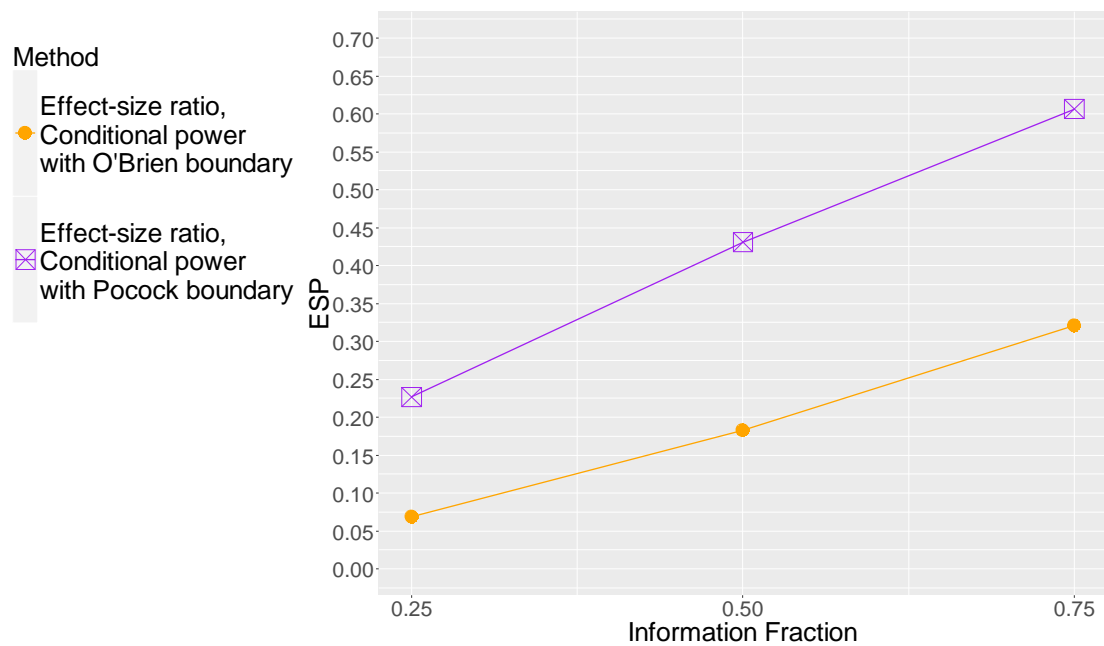


**Figure 7.52.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio.

#### **Fourth scenario of hazard ratio (i.e. 0.765)**



**Figure 7.53.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 1 for the fourth scenario of hazard ratio.

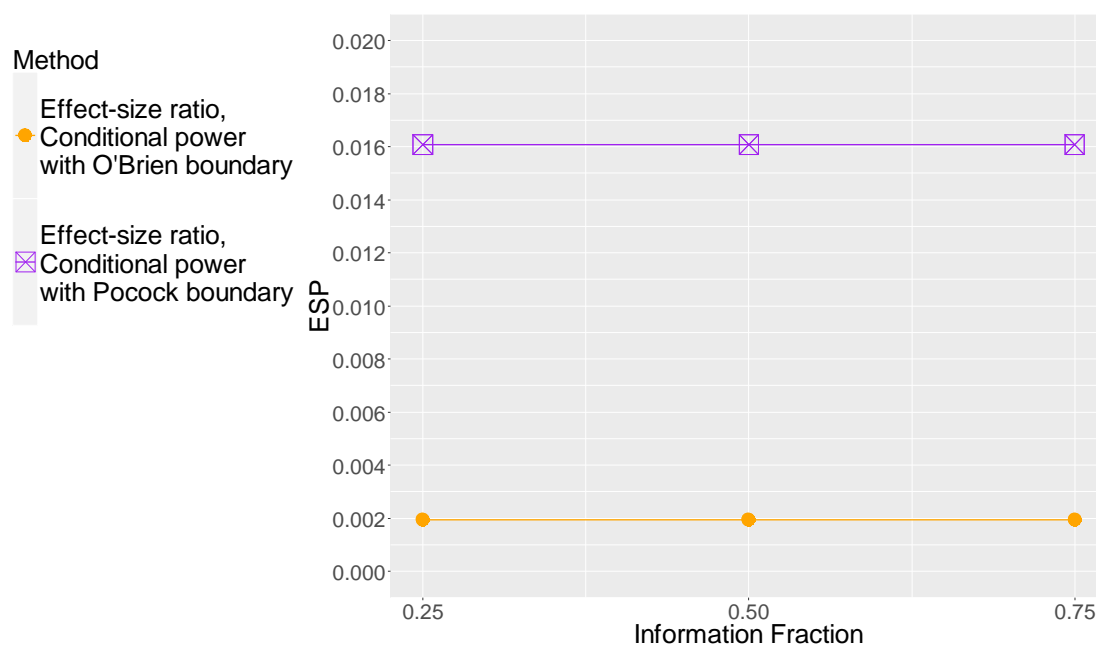


**Figure 7.54.** Efficacy stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio.

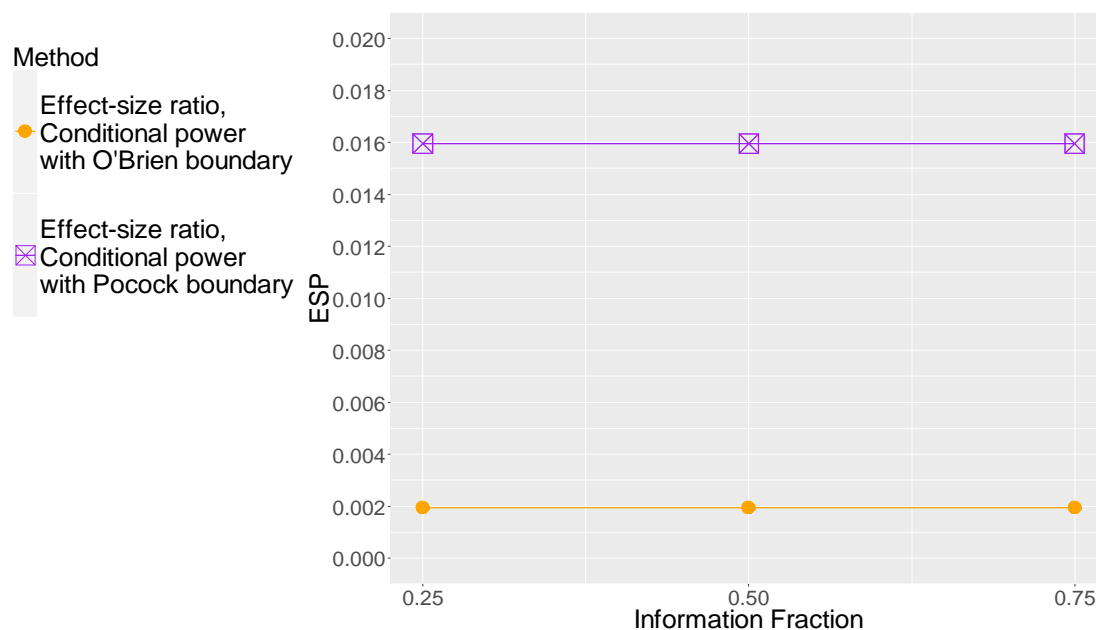
Similarly to the alternative hypothesis, in both simulation studies, under the null hypothesis, the efficacy stopping probabilities depend only on the efficacy stopping boundaries and not on the sample size adjustment methods. Figures 7.55 to 7.62 show that the efficacy stopping probabilities remain the same across the different percentages of information fraction for each type of stopping boundaries for all scenarios of hazard ratio. Higher values are achieved when the Pocock stopping boundaries are used compared to the O'Brien-Fleming efficacy boundaries.



### First scenario of hazard ratio (i.e. 0.746)

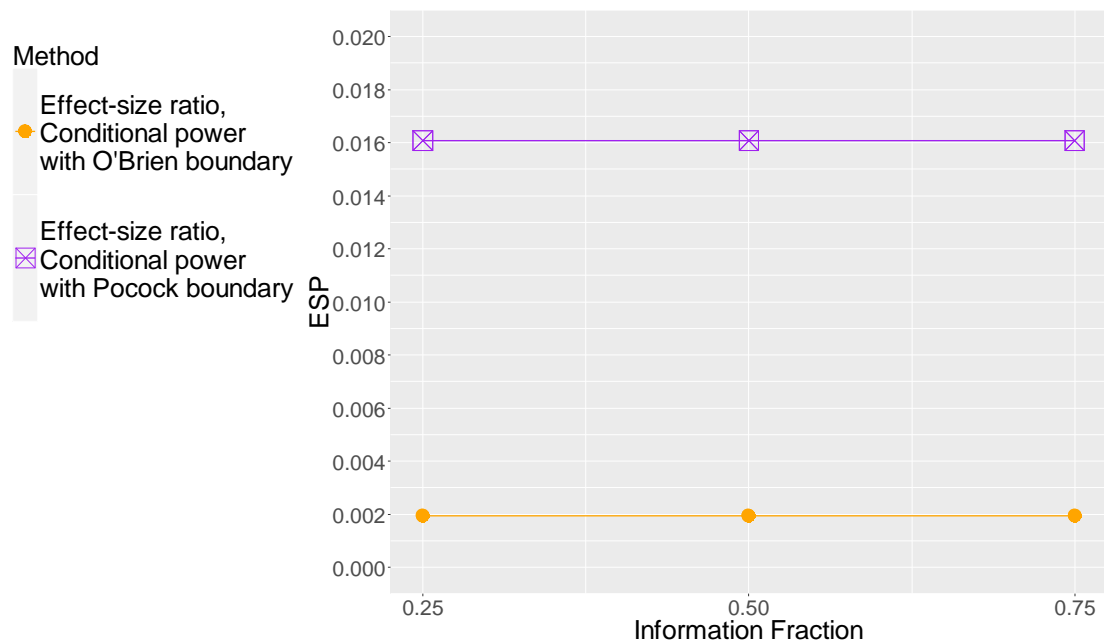


**Figure 7.55.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the first scenario of hazard ratio.

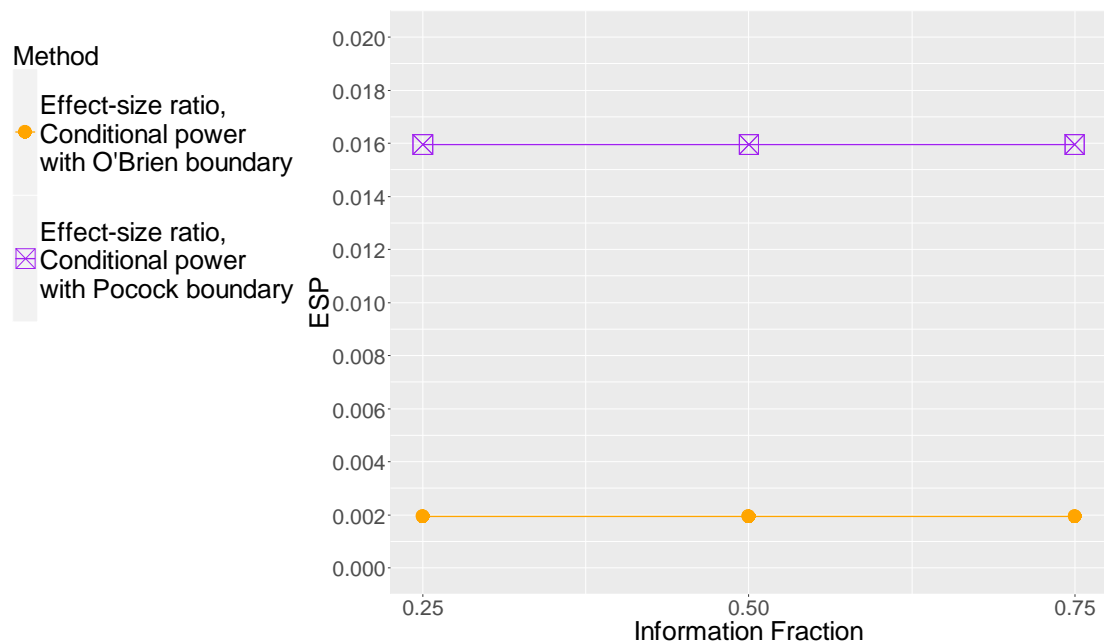


**Figure 7.56.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio.

### Second scenario of hazard ratio (i.e. 0.845)

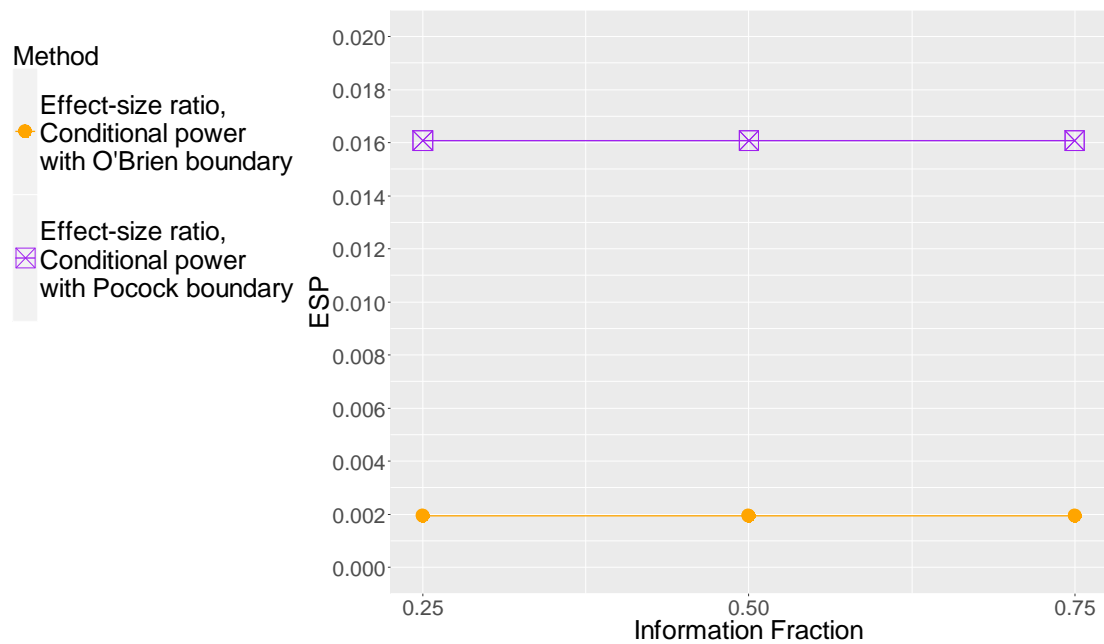


**Figure 7.57.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the second scenario of hazard ratio.

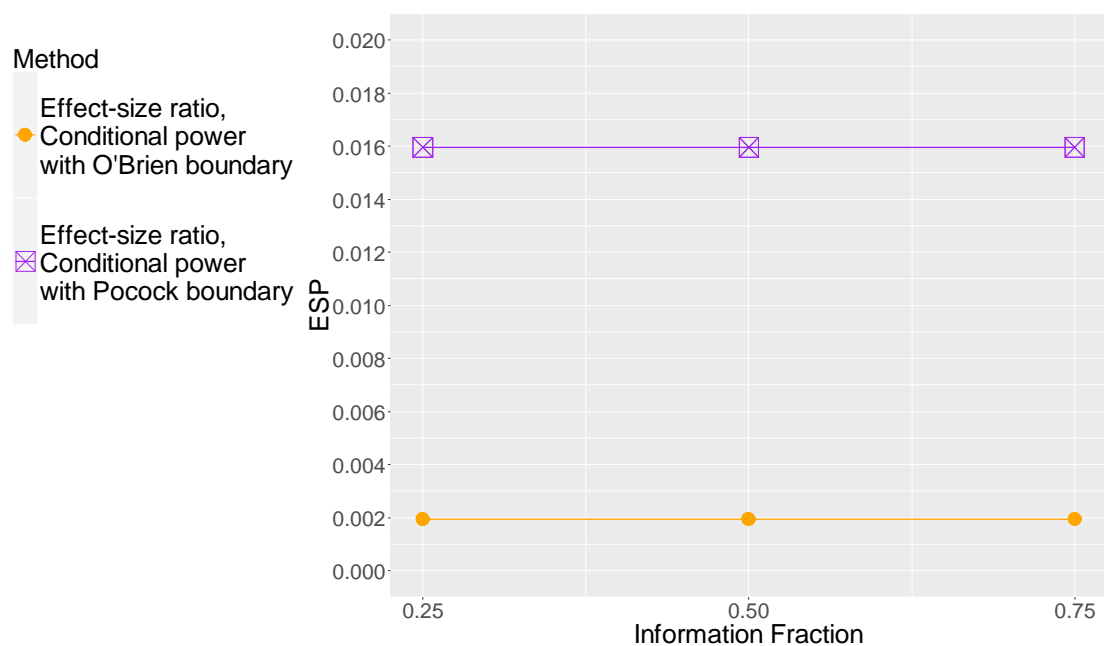


**Figure 7.58.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio.

### Third scenario of hazard ratio (i.e. 0.807)

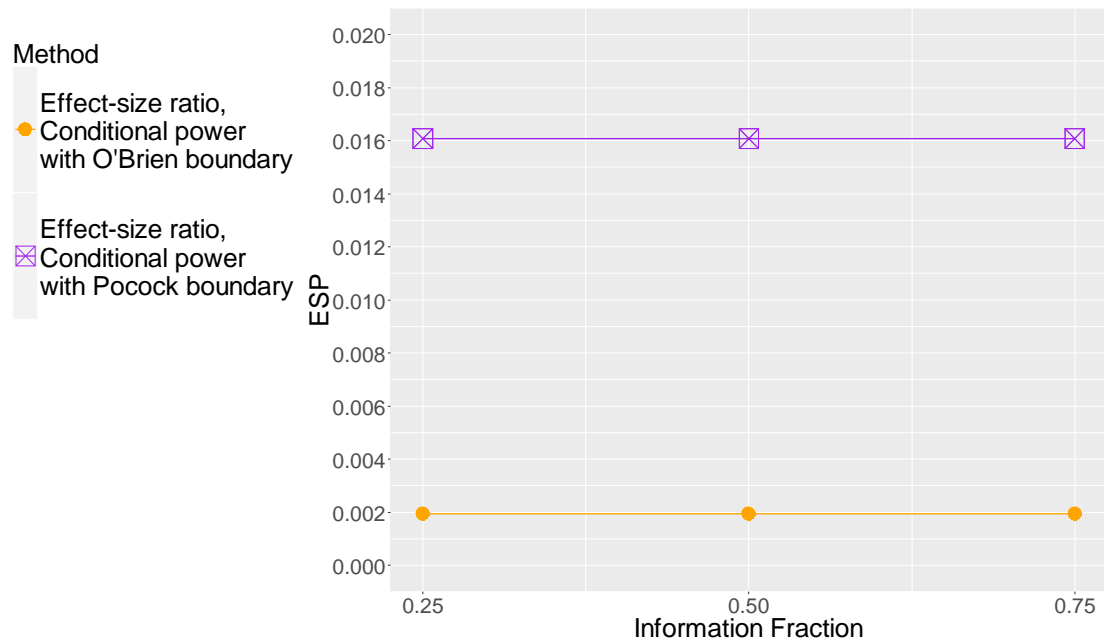


**Figure 7.59.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the third scenario of hazard ratio.

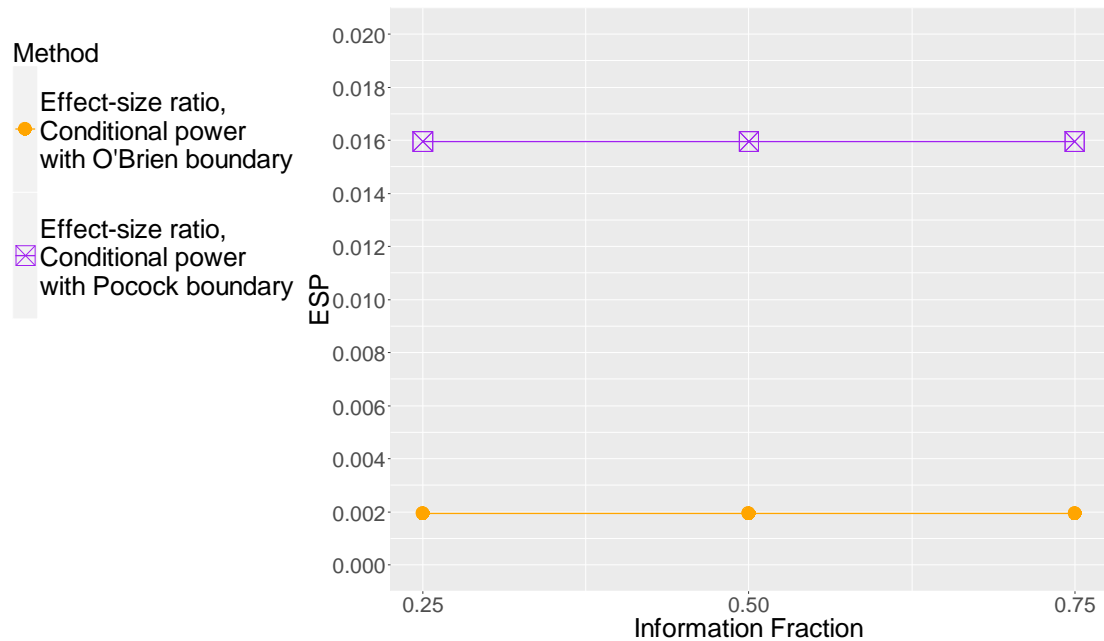


**Figure 7.60.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio.

#### Fourth scenario of hazard ratio (i.e. 0.765)



**Figure 7.61.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 1 for the fourth scenario of hazard ratio.



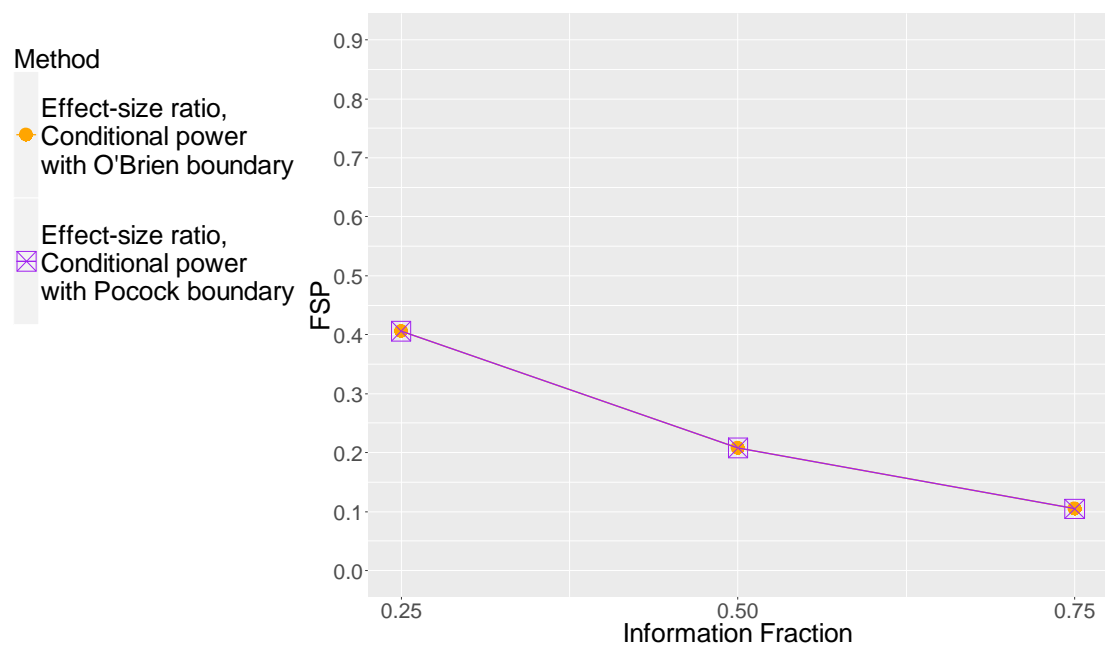
**Figure 7.62.** Efficacy stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio.

#### *7.5.2.5. Futility stopping probability*

In Simulation study 2, the futility stopping probability in all scenarios of hazard ratio depends only on the type of stopping boundaries and not on the method of

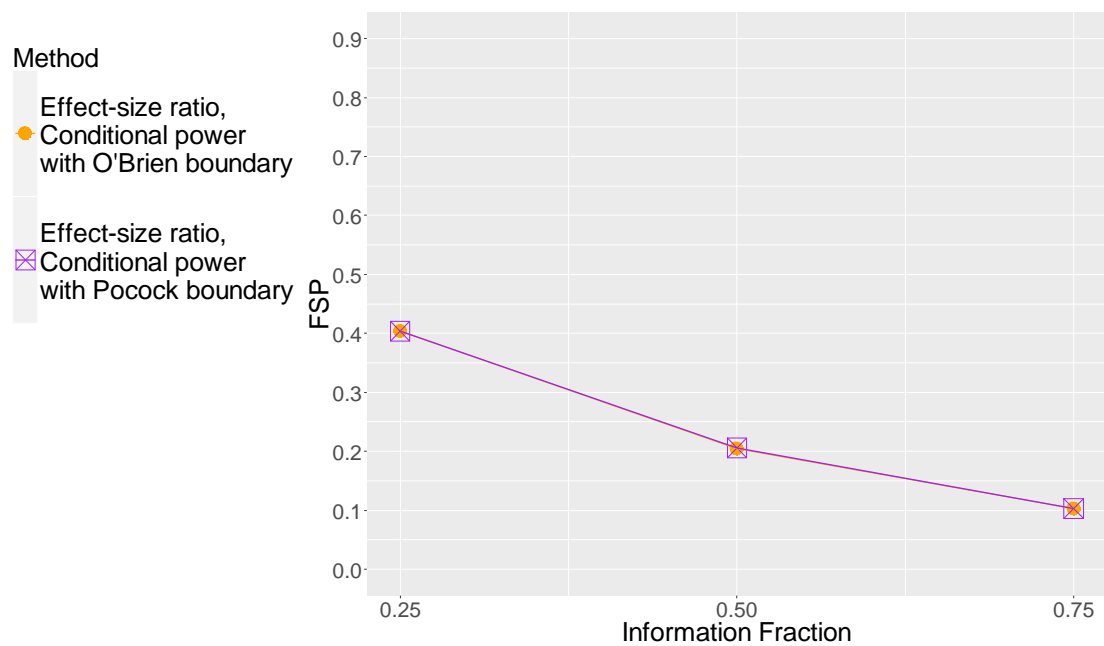
sample size recalculation. It decreases with the increase of information fraction under the alternative hypothesis (Figure 7.63 to 7.66) and it depends only on the type of stopping boundaries and not on the method of sample size recalculation. Hence, the smallest value of futility probability is obtained at 75% information fraction. The application of Pocock stopping boundaries results in slightly higher futility stopping probabilities compared to the use of the O'Brien-Fleming efficacy boundaries.

**First scenario of hazard ratio (i.e. 0.746)**



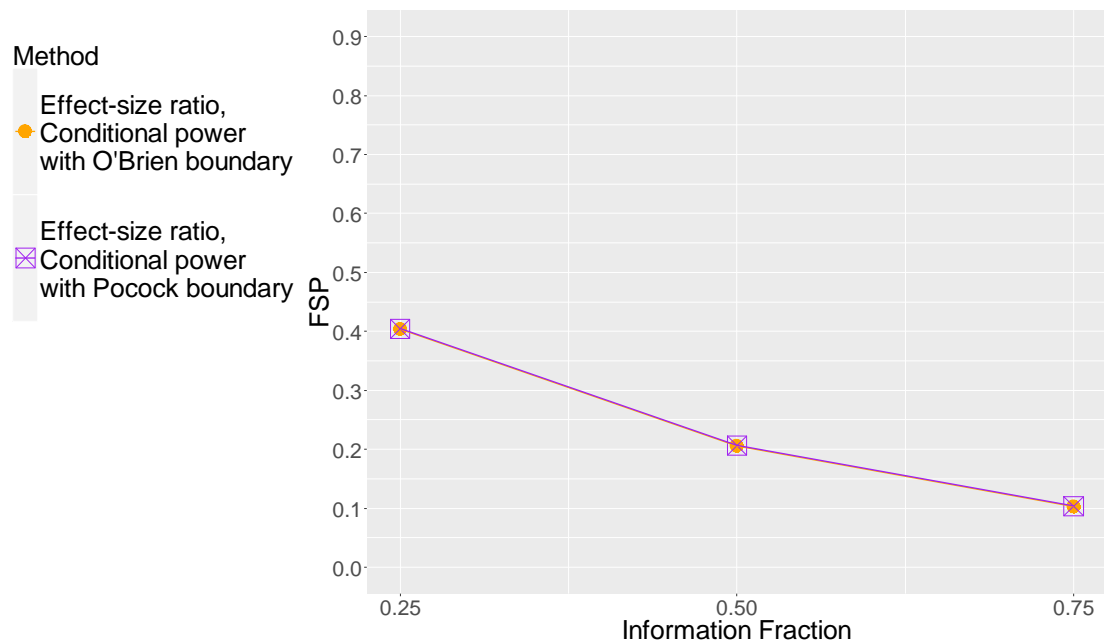
**Figure 7.63.** Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the first scenario of hazard ratio.

### Second scenario of hazard ratio (i.e. 0.845)



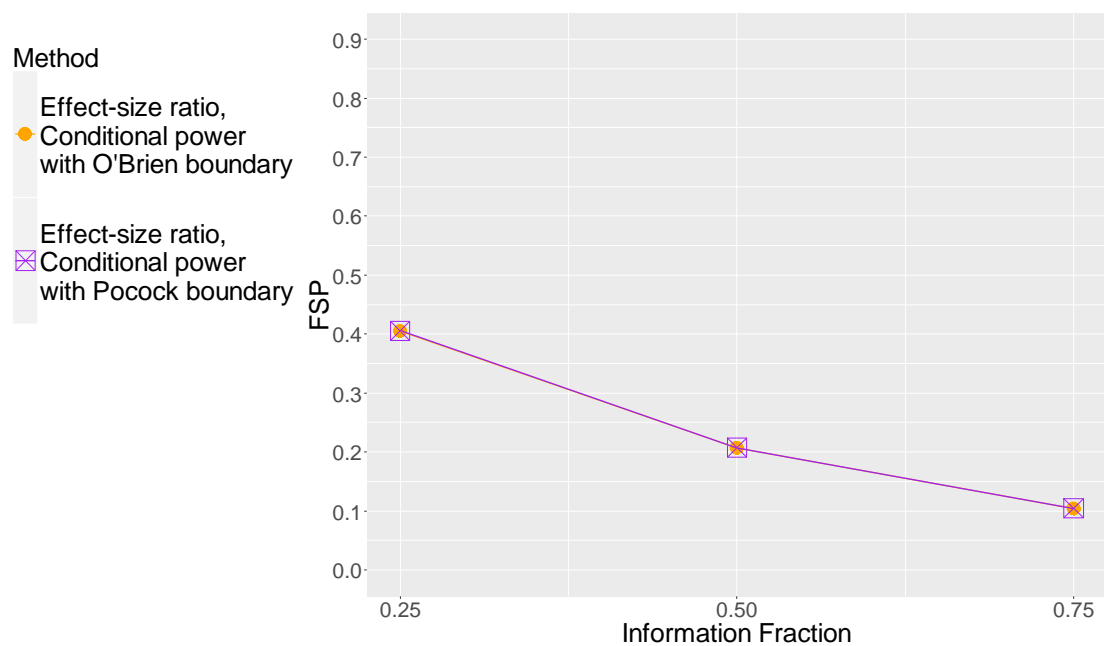
**Figure 7.64.** Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the second scenario of hazard ratio.

### Third scenario of hazard ratio (i.e. 0.807)



**Figure 7.65.** Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the third scenario of hazard ratio.

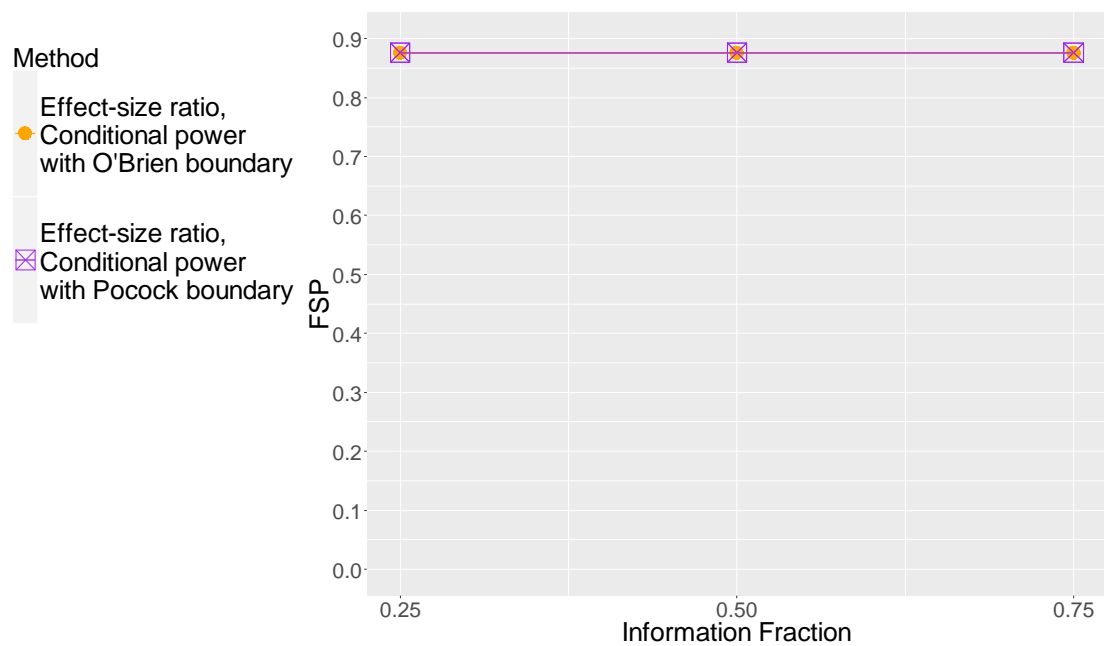
#### Fourth scenario of hazard ratio (i.e. 0.765)



**Figure 7.66.** Futility stopping probability versus the information fraction under the alternative hypothesis of Simulation study 2 for the fourth scenario of hazard ratio.

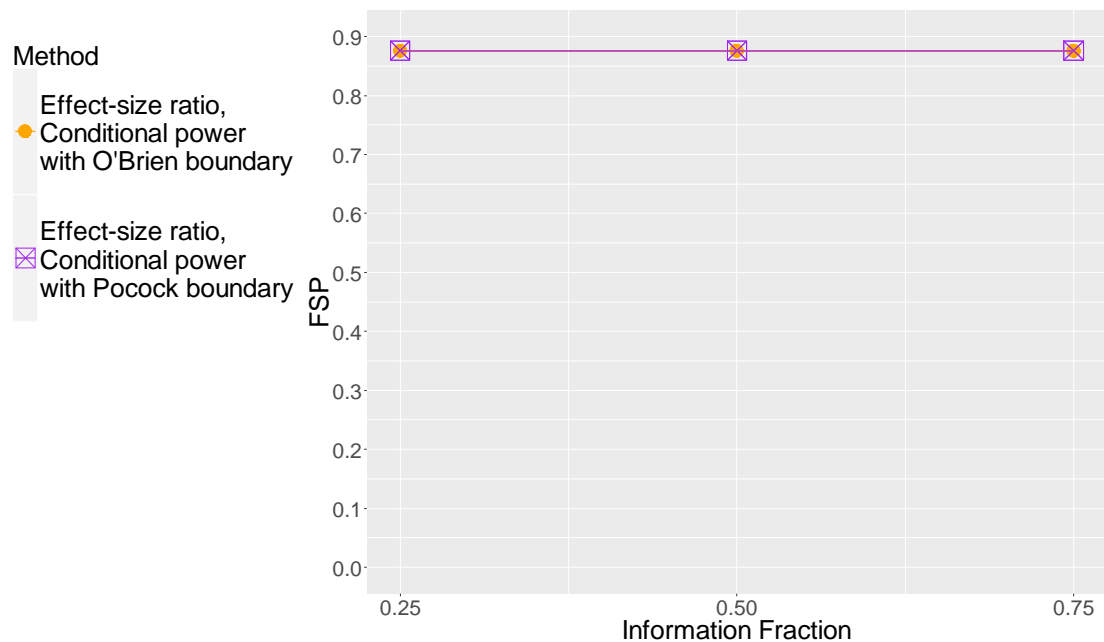
Similarly to the alternative hypothesis, the futility stopping probabilities in all scenarios of hazard ratios depend only on the efficacy stopping boundaries and not on the sample size adjustment methods under the null hypothesis in both simulation studies. Figures 7.67 to 7.70 show that the futility stopping probabilities remain the same across the different percentages of information fraction for each type of stopping boundaries. Similar values are achieved with each type of stopping boundaries.

### First scenario of hazard ratio (i.e. 0.746)



**Figure 7.67.** Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the first scenario of hazard ratio.

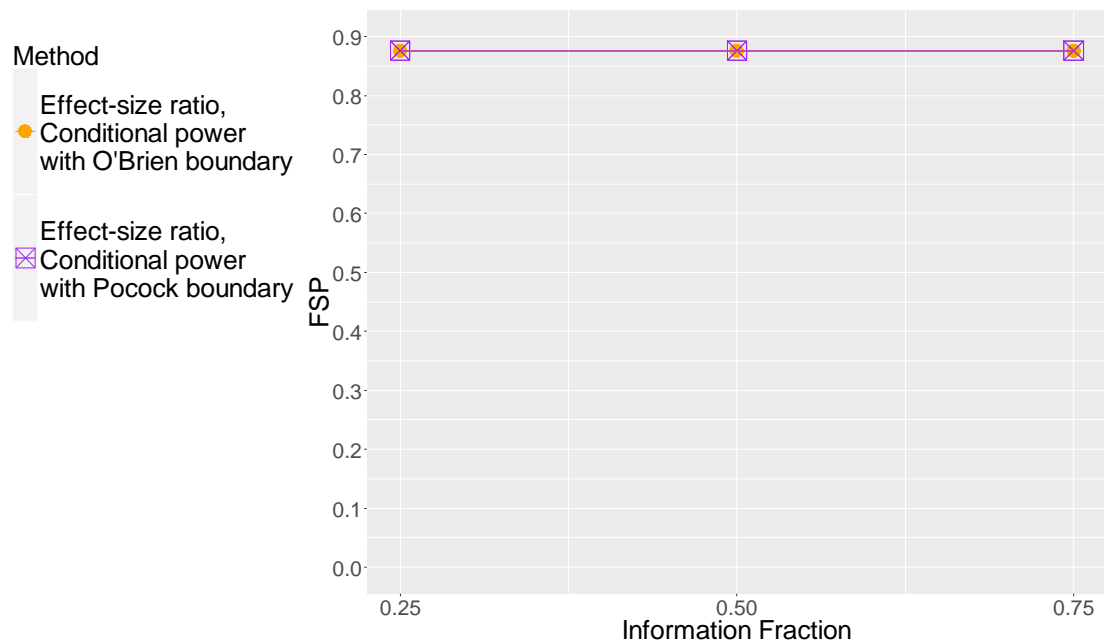
### Second scenario of hazard ratio (i.e. 0.845)



**Figure 7.68.** Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the second scenario of hazard ratio.

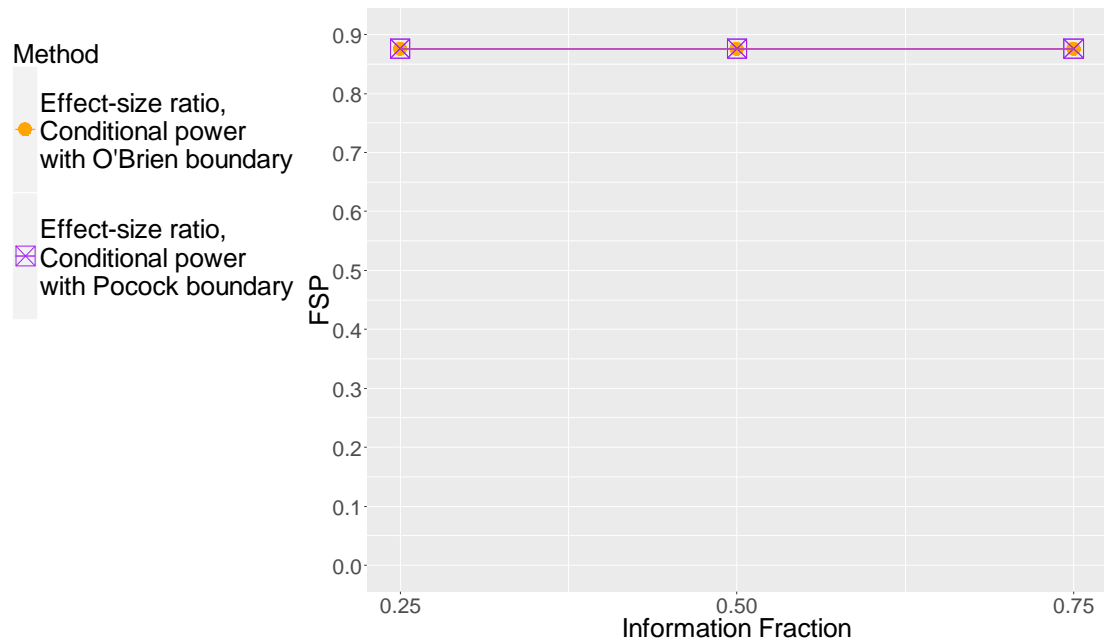


### Third scenario of hazard ratio (i.e. 0.807)



**Figure 7.69.** Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the third scenario of hazard ratio.

### Fourth scenario of hazard ratio (i.e. 0.765)



**Figure 7.70.** Futility stopping probability versus the information fraction under the null hypothesis of Simulation study 2 for the fourth scenario of hazard ratio.

## 7.6. Discussion

---

This chapter has focused on unblinded sample size reassessment. Generally, this approach is more controversial compared to the blinded method and has been characterized by the U.S. Food and Drug Administration (FDA) as a “less well understood approach” as it faces challenges with regard to operational bias [9]. In practice, several considerations should be taken into account (e.g. regulatory requirements, resources, efficiency etc.) before implementing these methods [10].

Our results show that there is a significant increase in the number of patients when the futility stopping boundary is not adopted, compared to the sample size which is needed for the non-adaptive design. The increase of the sample size is greater when the information fraction increases. When the futility stopping boundary is considered in our study, a reduced sample size can be seen from our results at all levels of information fraction. Furthermore, our results show that generally we can achieve greater power compared to the nominal level of power by using the sample size adjustment methods when allowing the trial to stop either for efficacy or futility.

It is important to note that at 25% information fraction, the futility boundary is too permissive and the trial is stopping too often. Hence, the futility boundary should be adjusted for the different information fractions.

Additionally, the current simulations assume that first stage patients are fully assessed before the interim analysis starts, and second stage patients are recruited afterwards. However, in practice, recruitment usually continues at an interim analysis, so this study could be extended to approach a more realistic scenario.

In general, sample size re-estimation is more difficult when a time-to-event outcome is used because when an interim analysis takes place, some patients who were recruited in stage 1, may not have an event by the time of analysis. Different approaches to this issue can be found in [11-13].

Different scenarios, methods for sample size recalculation and various options for stopping boundaries for efficacy and futility proposed to date can result in either satisfactory or poor statistical properties. One size does not fit all, thus, it is of utmost importance to study carefully through simulations the implication of each choice to the trial and investigate which option could be adjusted to a particular case.

In this chapter, the sample size re-estimation approach was applied to a biomarker-guided clinical trial design which was chosen as the optimal for STRONG trial presented in Chapter 6. This approach is similar to a regular randomized controlled trial, however, the difference is in the calculation of the required total number of patients which is based on the sample size formula of Reverse Marker-Based strategy design. In the next chapter (Chapter 8), we will describe various practical challenges of biomarker-guided trials, such as funding, ethical and regulatory issues, recruitment, monitoring samples and laboratories, biomarker assessment, data sharing, and resource that should be addressed when investigators conduct a biomarker-guided clinical trial.

## 7.7. References

---

1. Chang M. Adaptive Design Theory and Implementation Using SAS and R, Second Edition. 2nd ed. London: CRC Press; 2014.
2. Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics*. 1999; 55(4).
3. Antoniou M, Kolamunnage-Dona R, Jorgensen AL. Biomarker-Guided Non-Adaptive Trial Designs in Phase II and Phase III: A Methodological Review. *Journal of Personalized Medicine*. 2017; 7(1). doi: 10.3390/jpm7010001.
4. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 3rd ed. New York, NY: Springer; 2012.

5. DeMets DL, Lan K. Interim analysis: The alpha spending function approach. *Statistics in medicine*. 1994; 13(13-14):1341-52.
6. Kennedy R, Wang G, Cutter G, Schneider L. Effect of sample size re-estimation in adaptive clinical trials for Alzheimer's disease and mild cognitive impairment. *Alzheimer's & Dementia*. 2013; 9(4):P691. doi: 10.1016/j.jalz.2013.04.363.
7. Kiefer F, Jahn H, Tarnaske T, Helwig H, Briken P, Holzbach R, et al. Comparing and Combining Naltrexone and Acamprosate in Relapse Prevention of Alcoholism. *Archives of General Psychiatry*. 2003; 60(1). doi: 10.1001/archpsyc.60.1.92.
8. Lan K, DeMets D. Further Comments on the Alpha-Spending Function. *Statistics in Biosciences*. 2009; 1(1):95.
9. Gallo P, Anderson K, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, et al. Viewpoints on the FDA draft adaptive designs guidance from the PhRMA working group. 2010; 20(6). doi: 10.1080/10543406.2010.514452.
10. Pritchett YL, Menon S, Marchenko O, Antonijevic Z, Miller E, Sanchez-Kam M, et al. Sample Size Re-estimation Designs In Confirmatory Clinical Trials—Current State, Statistical Considerations, and Practical Guidance. *Statistics in Biopharmaceutical Research*. 2015; 7:4, 309-321. doi: 10.1080/19466315.2015.1098564.
11. Schäfer H, Müller HH. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in medicine*. 2001;20(24):3741–3751. doi: 10.1002/sim.1136.
12. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal*. 2006;48(4):714–729. doi: 10.1002/bimj.200510190.
13. Jahn-Eimermacher A, Ingel K. Adaptive trial design: A general methodology for censored time to event data. *Contemporary clinical trials*. 2009;30(2):171–177. doi: 10.1016/j.cct.2008.12.002.

## Chapter 8. Challenges in Practice

---

### 8.1. Introduction

---

In the previous chapter (Chapter 7) we performed simulation studies to explore the statistical and operating characteristics of an adaptive approach applied to the optimal design chosen in Chapter 6. In this chapter we investigate the challenges faced in practice when implementing a biomarker-guided trial, including those related to funding, ethical and regulatory issues, recruitment, monitoring samples and laboratories, biomarker assessment, data sharing, and resources. To identify and explore the key challenges arising, the author of this thesis, together with the leaders of the MRC Hubs for Trials Methodology Research Network's Stratified Medicine Working Group (SMWG) organized a workshop, 'Biomarker-guided trials: challenges in practice' and invited delegates with practical experience of conducting biomarker-guided trials from various disciplines including statisticians, trial managers, information systems specialists and clinicians. The workshop, held at the University of Liverpool in London Campus on 15th March 2017, was attended by 25 participants and the findings from the day form the basis for the content of this chapter.

In section 8.2 we describe some of the planned and ongoing trials the workshop delegates have been involved with, and in the remaining sections discuss practical challenges informed by their experiences working on these trials, together with some of my own reflections on those issues. Therefore, it is important to note that any statements made in this chapter are based on the delegates' personal experiences.

### 8.2. Examples of clinical trials

---

*i) The National Lung Matrix trial (ongoing trial) [1]:* This is a phase II non-randomized umbrella trial consisting of multiple single arm trials within one protocol. The aim of the trial is to investigate several new treatments hypothesized to

be of benefit to patients with advanced non-small cell lung cancer (NSCLC), and for whom surgery and radiotherapy are not deemed appropriate treatments.

The Matrix trial runs alongside the Cancer Research UK Stratified Medicine Programme (SMP2), where a next generation sequencing 28 gene panel test is used to assess the genetic profile of trial participants, which then determines which single arm trial (strata), and hence drug, they are assigned to. The trial adopts a Bayesian adaptive design with an interim analysis at 15 patients for each strata and final analysis of a target group of 30 patients per strata. The primary outcome is tumour shrinkage rate of at least 30%, which is deemed desirable.

*ii) Phase II trial of olaparib in patients with advanced castration resistant prostate cancer (TOPARP) (ongoing trial) [2]:* This is an open label, phase II, single arm adaptive trial for biomarker-driven selection based on response rate. It aims to evaluate the anti-tumour activity of the Poly (ADP-ribose) polymerase (PARP) inhibitor, olaparib, in metastatic castration resistant prostate cancer (mCRPC) and to identify molecular signatures for PARP inhibitor sensitivity with a pre-planned analysis to identify a biomarker-defined sensitive subgroup. First, unselected (i.e. without biomarker guided patient selection) mCRPC patients are all treated with olaparib and during the first stage of the design, if the response rate is high (i.e. > 50% responding) then further patients are recruited and a randomized placebo controlled clinical trial to evaluate the efficacy and safety of olaparib in these unselected mCRPC patients is undertaken. If the response rate is low (i.e. response rate < 10%), the trial is stopped. If in the intermediate range (10-50% responding), potential biomarkers of response are investigated and if a potential biomarker is identified, with those positive for the biomarker having a high response rate (>50%), the trial continues to the second stage with inclusion of the biomarker selected patients.

*iii) Adaptive multi-arm phase II trial of maintenance targeted therapy after chemotherapy in metastatic urothelial cancer (ATLANTIS) (ongoing) [3]:* This is an adaptive multi-arm randomized phase II trial which aims to explore whether

maintenance targeted therapy after chemotherapy, with treatment selection based on biomarker profile, delays time to progression and increases overall survival for patients with advanced urothelial cancer.

*iv) PRIMUS (ongoing) [4]:* This is an adaptive phase II trial, with biomarker evaluation integrated into the trial, which aims to assess the efficacy of FOLFOX-A (FOLFOX and nab-paclitaxel) when compared to AG (nab-paclitaxel and gemcitabine) in patients with metastatic pancreatic cancer, both in a biomarker-positive group and in biomarker-unselected patients).

*v) SALONICA (planned trial):* This is a stratified adaptive trial in ovarian cancer aiming not only to detect the key genomic determinants of response and resistance to neoadjuvant platinum-base chemotherapy in high-grade serous ovarian cancer but also to identify and validate putative biomarkers as well as test several novel drugs and corresponding putative biomarkers in women with poor response to neoadjuvant platinum chemotherapy through a phase II trial platform.

*vi) TASTER (planned trial):* This trial aims to identify predictors, at the point of diagnosis, of the efficacy of standard-of-care or some novel combination therapies in Chronic Myeloid Leukaemia (CML) patients who do not respond to tyrosine kinase inhibitor therapy. Both in vivo models of drug response and clinical data will be used to identify molecular signatures of stem cell resistance.

*vii) POETIC (Peri-Operative Endocrine Therapy for Individualizing Care) (ongoing trial) [5]:* This is a randomized, multicentre Phase III trial which aims to investigate whether having perioperative aromatase inhibitor (AI) therapy two weeks before and two weeks after surgery is more effective for postmenopausal women with ER+/PgR+ positive invasive breast cancer than having standard care alone. 4,476 patients were recruited from 130 UK centres with more than 2000 clinical staff. Patients received either AI therapy for 4 weeks (two weeks before and two weeks after surgery) or no AI therapy.

*viii) FOCUS4 trial (ongoing trial) [6]:* This is an umbrella clinical trial consisting of parallel, molecularly stratified randomized comparisons in patients with metastatic colorectal cancer (mCRC). Patients with newly diagnosed mCRC are registered into the trial and commence their standard first line chemotherapy which typically lasts for approximately 16 weeks. During this time, a sample of their tumour is sent away to one of two dedicated FOCUS4 laboratories who perform genomic and molecular tests on the tumour. This enables stratification of the patients into one of a number of pre-specified molecular subgroups (called cohorts). Patients are then offered entry into a randomized trial (called comparison) testing a specific targeted therapy for their subtype of cancer. All these comparisons are randomized and controlled and wherever possible use a placebo in the control group.

### 8.3. Challenges

---

#### 8.3.1. Funding issues

---

Although funders appear to be enthusiastic about supporting biomarker-guided trials, due to their complexity the resources required to deliver them are typically substantially higher than for trials with more simple designs. This needs to be factored into the decision making process when deciding on whether to fund, recognising that despite increased costs the trial may well be more efficient in demonstrating patient benefit. When considering the additional resources required, the increased administrative burden on staff should not be forgotten, for instance those arising with umbrella type designs where all the necessary paperwork has to be repeated for each part of the trial. Funders also often struggle with the monitoring of biomarker-guided trials, particularly those with an adaptive design, and are often uncertain and inconsistent in how they handle amendments to the design. It is typical for an amendment that is cost neutral to be approved quickly without additional approvals, but if the amendment is likely to cost money (e.g. the addition of a new trial arm) then it has to go through the more classic route of peer-review and funding approval. It is possible that research groups with experience of running such trials



could work together with funders to help inform and advise them on the implications of using such designs for their funding streams.

To avoid the same peer review process being triggered with each addition of trial arms, funders may expect applicants to provide estimated details of these potential additions at the outset to allow them to earmark the foreseeable additional budget and provide approval in principle. So, a researcher submitting an application for an umbrella trial, for example to include initial arms A to D, would be required to also estimate how much it would cost to make changes E, F, and G at time points X, Y and Z. However, this can be difficult as it requires knowledge not only of the approximate size of the cohorts to be added (or indeed removed) within those changes, but also the time point at which they will be added and the approximated end date. In addition, including additional forecasted costs could easily make a trial very unattractive to funders due to its expense, for example projected total costs for a large trial could use up the entire budget for a particular funding call. Funders will always be limited by the pot of money available within a particular fiscal period and will be faced with many competing funding requests, many of which will have simpler and easier to understand designs with more transparent budgets.

To convince funders, it may be that some of the currently ongoing trials need to be completed to understand whether they represent good value for money, although this could be misleading since it is widely felt that many of the trials are significantly underfunded as they currently stand. Quite often, it is the Clinical Trial Units (CTU) costs (e.g. trial management, trial monitoring, trial statistician) that are compromised because they are more malleable than other costs. It is possible that academia and funding bodies could work together in order to co-create funding solutions. For example, funding bodies with a long-term research commitment in a particular disease area (e.g. Cancer Research UK) could provide the essential long-term core funding for such a trial (e.g. CTU costs) by setting aside a proportion of their annual budget indefinitely, while other funding bodies with a shorter-term vision (e.g. pharmaceutical companies or smaller charities) could provide add-on funds associated to a particular treatment arm of their interest (i.e. additional

infrastructure and personnel costs). A funding structure of this type would not only provide sufficient funds for successful trial delivery but could also allow for more transparent and efficient allocation and monitoring of resources while creating cost saving opportunities through synergies.

There has also been some confusion about who should fund the additional biomarker tests within a trial. Funders have previously suggested that this is a National Health Service (NHS) associated cost since it is used to direct treatment, however the test is often not available on the NHS, and thus they have refused to fund it. The situation may be slowly changing, however, since we are moving into an era where more biomarker tests are going to be routinely undertaken in practice.

Finally, an additional funding issue is related to whether the trial uses previously untested biomarkers or more established and validated ones; the former may incur additional costs for the development, validation and standardization of appropriate tests, delays in the expected start date and recruitment problems due to poor quality of biomarker assessment results.

Overall, we believe that detailed planning and clear communication between researchers and funders are vitally important to ensure that future trials can be fairly considered and appropriately funded. There is also room for education, with those with practical experience of such trials sharing their knowledge with funding bodies. Trialists and funders alike may feel overwhelmed by these trials but in fact they are not as complicated as is often believed. Broken down into separate arms they can be considered as individual trials being controlled by an overarching research team and with some additional biomarker analyses. If they are broken down in this way and communicated effectively, then they should not be feared.

### 8.3.2. Ethical and Regulatory Issues

---

A key issue here is the confusion that sometimes exists within regulatory bodies about the characteristics of a biomarker-guided trial, for example, understanding the difference between an umbrella and a MAMS trial and how these are different from

a conventional trial. For instance, there is an expectation that when adding a new Investigational Medicinal Product (IMP) to an umbrella trial there should also be a new CTA (Clinical Trial Authorization), which is clearly not the case. It is important for regulators to appreciate that adaptations to the design are part of the original regulatory approval submission. There is also often the belief that from a commercial perspective the trial will be testing, developing and marketing a companion diagnostic alongside the therapeutic which is usually not the case.

Although there is general consensus that research ethics committees are very positive and receptive to these types of trials, there are many ongoing administrative issues that would benefit from being addressed. Whilst an ethics committee might give their overall ethical approval at the very start of a trial, it is often not clear how the addition of new arms will be approved at a later date. Depending on the local practice, such amendments may not be reviewed, discussed and approved by a sub-committee or may even come through simply as a chairman's action. Consequently, the trial documents are perhaps not checked as thoroughly as the original application and the amendments may not be scrutinised in sufficient detail. In addition, there is inconsistency in terms of what documentation ethics committees request to approve such an amendment, with some requesting a new submission and others seeking a major amendment. It is important that discussions are held with the HRA to ensure that their administrative systems, paperwork, and their version control is adapted to deal with these types of amendments adequately. Researchers with experience of running such trials would be well placed to advise in this regard.

There are similar administrative issues that need addressing with the Medicines and Healthcare products Regulatory Agency (MHRA), and similar discussions could be had with them. For example, the name of a trial's CTA is given in accordance with the initial treatment arms included in the trial; however, these may not be part of the trial after a while which can lead to confusion in terms of terminology.

From the perspective of patients, some challenges relating to the informed consent process may arise. There are examples of having to consent patients into the trial on the same day of diagnosis, which clearly requires a lot of sensitivity and both careful and appropriate communication. There is also an issue related to the fact that biomarker screening might fail requiring a second biopsy. Obtaining a second biopsy can be difficult because patients are often not well enough. Having a trial option for non-stratified patients (including those with failed biopsies) can be a good idea, particularly if biomarker screening is invasive or has a high failure rate.

Effective communication to patients is also fundamental to ensure that there is a clear understanding of why these sorts of biomarker trials are undertaken, as whilst they are often about targeting treatments to patients who are believed to have the best chance of benefitting, they can also be about trying to avoid treatments in patients who don't need them. This will aid acceptance by those being denied a treatment due to their biomarker profile. Trials would also benefit from improving how personalized medicine is described to the public, since this can be misleading. Whilst on the surface personalizing treatment may sound like the perfect solution, patient expectations need to be managed. It should not be communicated as an approach to ensure that a treatment will definitely work in a patient with given biomarker status, but rather is an approach that will mean it is more likely to work.

An additional ethical challenge, this time more so from the researchers' perspective is that since patients' samples are often being genotyped, there is always the risk that susceptibility to certain diseases are uncovered. Whilst this issue can be covered in the informed consent process from the patient's perspective, it can pose a moral dilemma to those involved in conducting the trial. Additionally, from the patients' perspective, they can often assume that having certain mutations in their tumour means an increased risk of disease in relatives. Hence, careful communication is again needed in order to clarify the difference between mutations in a tumour and germline mutations (i.e. hereditary mutations passed on from parents to offspring).

In summary, several ethical and regulatory challenges can arise ranging from a lack of understanding about administrative procedures to issues relating to communications with patients. It is essential that accurate information about biomarker-guided trials are communicated to all relevant stakeholders so that they are educated on the characteristics and advantages of such trials.

### 8.3.3. Recruitment

---

Uncertainty in recruitment rates, especially in trials which include rare biomarker groups can be a big concern. The prediction of recruitment rate into umbrella trials can be difficult as several factors need to be considered: 1) the estimated prevalence of each biomarker, which might not be accurately known at the design stage and in the case of trials that evaluate multiple biomarkers might also be affected by overlapping groups; 2) the failure rate of lab diagnostic biopsies in the technology hubs; and 3) additional factors related to the consent rate. The difficulties are compounded by the fact that funders and sponsors regularly question whether the reached recruitment rate is close to the projected. If not, they may require recalculations and protocol amendments which can often be more complex for biomarker-guided trials, in particular umbrella trials, than for a traditional trial. Hence, new methodology for predicting the rate of recruitment may be needed for these trials.

Recruitment issues can be patient related or researcher related. From the patients' perspective, they may not want to join the trial for various reasons, for example, they may be weary after the first line treatment, their disease may have progressed, or they may simply not be interested in the new drug and would like to take a break from treatment. In addition, having complex tissue sampling (mandatory fresh biopsies) is always a challenge for recruitment since patients are generally less interested.

From the researchers' perspective, slow set up of comparisons may occur leading to sites losing their enthusiasm which can have an impact on the recruitment

of patients. In turn, this may affect the motivation of commercial partners to get involved. Further problems can arise when it is difficult to predict recruitment timelines. Another barrier to timely recruitment may be a high screen failure rate, requiring an increase in screening activity at site with further cost implications. Additionally, it can take longer than the expected timeframe to deliver biomarker results to centres, which can have an impact on accrual.

In addition, the dropout rate from trials can be significant, particularly where trials involve very ill patients with rapid deterioration. Long waiting times for the genetic profiling can result in patients not being well enough to participate by the time their biomarker status was identified. This issue is often compounded further due to the length of time taken to test a patient's sample meaning that it is not uncommon for a patient to have died before the results are available. Even if they are still alive, the patients may have deteriorated and decide they no longer want to be involved in the trial. Risk of dropout is further increased since once someone has been randomized the workup for whichever treatment they are stratified to can be a very involved process, and the patient may decide to take the simpler option of not taking part in the trial. Also, receiving an extra novel drug may require travel to a further location and those in a terminal illness or with advanced disease may not wish to do so, and would rather try to enjoy their remaining months. Although it is impossible to predict what these patients will decide, their likelihood of dropping out can be reduced by ensuring rapid turnaround times for biomarker test results, and this should be encouraged.

To summarize, given the multiple factors impacting how quickly patients will be identified, recruited and retained in a biomarker-guided trials, estimating an accurate rate of recruitment will always be difficult. Further, even if patient accrual and retention is good, research staff responsible for recruiting patients may be less efficient if delays occur in setting up sites. It is suggested, therefore, that well-designed pilot and feasibility studies are undertaken prior to trial commencement to ensure a more accurate understanding of recruitment rate as well as a smoother and

more rapid process of site set-up. Laboratories should also be sufficiently equipped and efficient to deal with rapid biomarker analysis turnaround.

#### 8.3.4. Monitoring samples and labs

---

Providing a lab is accredited it is expected that good internal audit trails are in place, however logistical problems can occur in the transfer of results from the lab to the CTU. Often, data are not sent on an individual patient basis, but rather in batches of hundreds, or thousands at a time, so it is important to agree on procedures for transferring these data accurately in order that mix-ups do not occur. Problems can arise when lab staff are not always GCP (Good Clinical Practice) trained, and there have been examples of data confidentiality being breached when staff have not appreciated that the data were clinical trial data which were therefore required to remain confidential until the trial had been reported. Therefore, it is recommended that there should be a better understanding of GCP requirements within labs.

Monitoring patient samples requires a significant amount of work and coordination. For example, a first sample may be received and there might be insufficient tumour, meaning that another sample has to be requested. A full audit trail is therefore required (accurate and traceable matching of patients and tissue samples) to ensure that the correct value ends up in the analysis. A large amount of data cleaning is also typically required.

In terms of the handling and tracking of samples, local research nurses, pathologists, lab staff as well as the CTU should be involved. Further difficulties arise if the tissue obtained has inadequate tumour tissue or is not viable and further requests for samples need to be made back to the original hospital pathology departments.

To ensure optimal efficiency it is recommended that lots of samples are batched up to be sent all at once instead of using additional resources on several small runs. However, this can often lead to problems with lower than anticipated recruitment

leading to further delays whilst labs wait for enough samples to justify running a batch.

Another challenge associated with biomarker analysis is that science is advancing so quickly, things can change dramatically after the funding application stage with many additional opportunities arising. It is recommended that a separate lab manual is used outside the protocol in order to minimize any associated protocol amendments.

In terms of ensuring completeness and quality of tissue samples received, communication and collaboration between clinicians and laboratory staff should be strengthened to ensure that the samples are taken, stored and sent off in accordance with the protocol. In addition, the CTU's central trial monitoring capabilities should be utilized to ensure efficient sample tracking. In our view, strong collaboration between the CTU and the laboratory staff is much needed as the success of a biomarker-guided trial depends heavily on the accurate and timely delivery of lab results.

#### 8.3.5. Biomarker assessment

---

The setup of SMP2 is a major undertaking for biomarker trials, especially for these umbrella trials, and the trial's success relies on SMP2 being successful. Lung biopsies from advanced cancer patients can be really challenging, and this may apply to other diseases too.

One major challenge during biomarker assessment is that we are often dealing with small samples which can be heterogeneous and a mix of normal tissue with tumor tissue. This results in less confidence that a mutation, which is in fact present, will be detected, and therefore less confidence in the randomization procedure since it is directly linked to this assessment. Biomarker misclassification issues therefore represent a major challenge within biomarker-stratified trials, and should be addressed. The analytical validity of a biomarker in terms of sensitivity and specificity is a challenging but very important issue and understanding the accuracy



of an assay is a necessary consideration. If a sample fails completely, it is easy to class it as failed; if there is a partial fail, for example if 8 genes are tested and 6 of them pass completely but one of them passes 90% and another one fails 50%, this represents a difficult result to handle and it can be difficult to randomize a patient based upon such a result.

#### 8.3.6. Data sharing issues

---

When a pharmaceutical company is involved in a trial, alongside the clinical study report it is expected that the company will also request the trial data to be shared with them at the end of the trial, within a data sharing framework. However, there is sometimes pressure for the data to be shared in real time or at least at periodic intervals (e.g., to guide business decisions) throughout the time the trial is open. Current consensus suggests that this is not a good idea for phase III trials but in a phase II trial (particularly a non-randomized one) there are varied opinions as to its merits (especially when considering a response rate endpoint). One argument against this type of data sharing during the trial is that historically, if you questioned why a phase II trial had failed, one reason was that the clinicians or chief investigators were too selective to pick their patients when they had a fixed threshold of responders to reach to call it a success (e.g., picking patients more likely to respond creating a distorted cohort of patients in the latter part of the trial). Sharing data during the trial could result in these types of situations arising again. An argument for data sharing during the trial from the companies' perspective is that each of these patients are individuals and they will learn from the genomic profiles of those patients who the real responders are to their targeted agents. They want the ability to know this information effectively in real time so they can pool their data with the data from any other study that they are working on worldwide, so that they can learn and make a business case for the future development of their pharmaceutical product.

The direct involvement of pharmaceutical companies can prove challenging in general since, due to their financial investment, they will likely be keen for trial data to be shared with them on an ongoing basis. Further, whilst decisions in terms of the

closure of strata are the responsibility of the trial steering committee, pharmaceutical companies may wish to be heavily involved in the decision making process.

Data sharing requests from pharmaceutical companies are likely to be common in biomarker guided trials as the sharing of clinical data could enhance the medical research. However, differing viewpoints in terms of how and when data should be shared can be particularly challenging for the trial management team. To ensure that good relations are maintained with all interested parties, it is recommended that a clear data sharing policy and common data standards are developed and agreed at the beginning of the trial.

#### 8.3.7. Resources

---

In terms of CTU management, ensuring the availability of appropriate resources is a challenge. Biomarker-guided trials require a large number of personnel, a lot of effort, and a lot of money. Information Technology (IT) support is frequently underestimated and vital for many types of biomarker-guided trials. Data scientists serve a vital function and they need to be adequately costed into any grant application. For perpetual trials, databases and case report forms (CRFs) are important to get right as they could be used for a long time. The complexity of the CRF is another challenge, since the data required often vary between strata. Hence, there is a need to have several different CRFs and therefore the IT requirements are equivalent to several separate trials with an additional need for a more sophisticated database structure. Protocol amendments lead to additional problems due to the fact that for just one amendment (e.g. an additional medical assessment), all CRFs require modification.

Furthermore, administrative support for things such as preparing site packs is often underestimated, and the need for collaboration between a CTU and biomarker labs put further pressure on resources. A huge amount of biomarker expertise is required, which is not always available within a trials unit. Other challenges associated with their collaboration relate to lab agreements (e.g., impact on data

sharing) and the processes for tracking, blinding and pseudo-anonymization of samples.

More complex work is also needed when adding new comparisons. Several issues need to be considered at that time; a new trial needs to be set-up, including protocol and CRFs development, database development, setting up of contracts, drugs supply etc. while existing comparisons are already running, making the implementation of amendments at centres very challenging.

To summarize, the resources required for efficient management of a biomarker-guided trial should not be under-estimated and CTUs need to ensure that they are prepared in particular for the administrative burdens that come with such trials, and cost them into any funding applications.

#### 8.4. Recommendations to overcome practical challenges

---

- Define the objectives of the study and consider hypotheses carefully. Clinical trial team should decide whether they should employ a statistical test to prove that a treatment is superior to another (i.e. superiority trial) or that two treatments are not too different in characteristics (i.e. equivalence trial) or that a treatment is not worse than another (i.e. non-inferiority trial). Additionally, the significance level of a test (i.e. type I error) as well as the type II error should be pre-specified according to the level of evidence being sought (Phase II exploratory or Phase III confirmatory trial).
- Use BiGTeD to explore various designs according to hypothesis.
- Construct research questions which will result in the right choice of the study design. These questions can be developed by using the information related to the utility of biomarker-guided trial designs provided in Chapter 2 and 3 of this thesis as well as BiGTeD. An example of research questions with their answers can be

found in Chapter 6. For instance, questions could be asked to check whether: the evaluation of treatment effect would be performed only in biomarker-positive subgroup or not, the required sample size would be large due to low prevalence of biomarker, several treatments and biomarker levels exist, testing a clinical strategy would be of interest, there is uncertainty regarding the magnitude of the treatment effect that would suggest the use of an adaptive strategy, etc.

- Operating characteristics such as sample size, study power, trial duration of the clinical trial should be investigated through different simulations approaches (see Chapter 5, 7).
- Consider challenges faced in real practice which are presented in the current chapter. Some of these challenges are listed below:
  - Funding issues due to higher resources required for these designs compared to non-biomarker-guided trial designs, lack of funders' knowledge on the implications of using such designs, etc.
  - Ethical and regulatory issues mostly related to lack of understanding about administrative procedures to issues relating to communications with patients.
  - Recruitment issues regarding the estimation of an accurate rate of recruitment, delays in setting up sites, ill-equipped and inefficient laboratories.
  - Issues arise in the monitoring of samples and labs when lab staff are not GCP trained, communication and collaboration between clinicians and laboratory staff is not effective, etc.
  - Small samples which can be heterogeneous as well as the analytical validity of a biomarker in terms of sensitivity and specificity.
  - Challenging issues when pharmaceutical companies request sharing of clinical data.

- Underestimation of IT and administrative support, complex databases and CRFs.

## 8.5. Discussion

---

Here, informed by the workshop ‘Biomarker-guided trials: challenges in practice’ several practical challenges in conducting biomarker-guided trials have been considered. Although many of the challenges discussed relate to very large and complex biomarker-guided trials such as umbrella trials, similar challenges appear in biomarker-guided clinical trials more generally.

Despite the aforementioned challenges, the biomarker-guided trials discussed within this chapter represent hugely successful research projects using novel designs which will hopefully inform future trials.

To conclude, benefits of adopting biomarker-guided trial designs can arise, despite several teething problems resulting from using such novel methodologies. However, the significant investments required to successfully conduct such trials should not be underestimated, and it is imperative that the practical challenges they bring for clinicians, laboratories, regulators, academia, industry and patients as outlined above should be acknowledged and addressed at the outset. As the need for trials in stratified medicine increases, however, it is anticipated that through experience stakeholders will become more familiar with the designs, and the procedures involved in conducting and managing them will evolve and adapt accordingly. It is important therefore that the knowledge gained by those with experience of biomarker-guided trials is communicated to the wider research community such that all stakeholders are educated about the complex issues associated with biomarker-guided clinical trials and recommendations on how they may be overcome.

The final chapter (Chapter 9) gives a general overview of the topics discussed in this thesis and gives some future directions.

## 8.6. References

---

1. National Lung Matrix trial. Available online: <http://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-trial-looking-at-different-drugs-for-non-small-cell-lung-cancer-national-lung-matrix-trial>.
2. Phase II trial of olaparib in patients with advanced castration resistant prostate cancer (TOPARP). Available online: <http://www.icr.ac.uk/our-research/our-research-centres/clinical-trials-and-statistics-unit/clinical-trials/toparp>.
3. ATLANTIS: An adaptive multi-arm phase II trial of maintenance targeted therapy after chemotherapy in metastatic urothelial cancer. Available online: <http://www.crukctuglasgow.org/eng.php?pid=atlantis>.
4. PRIMUS 001. Available online: <http://www.crukctuglasgow.org/eng.php?pid=primus001>.
5. Trial of perioperative endocrine therapy - individualising care (POETIC). Available online: [http://www.icr.ac.uk/our-research/our-research-centres/clinical-trials-and-statistics-unit/clinical-trials/poetic\\_trial](http://www.icr.ac.uk/our-research/our-research-centres/clinical-trials-and-statistics-unit/clinical-trials/poetic_trial).
6. FOCUS4. Available online: <http://www.focus4trial.org/>.

## Chapter 9. Conclusions and Future Research

---

Recent advances in genomics and the heterogeneous nature of diseases and response to their treatment has led to an increasing interest in personalized medicine which aims to select the treatment approach best suited to a patient and promises to accelerate and improve drug development. Clinical trials are essential research tools for testing the safety and efficacy of treatments and can explore whether a treatment works well for patients and also which treatment is most effective for a particular subgroup of patients. Today, biomarkers are becoming an integral part of clinical trials as they are considered key tools in the identification of patient subgroups most likely to benefit or most likely to suffer adverse reactions from a given treatment [1, 2]. Hence, so-called biomarker-guided trial designs are pivotal in advancing the field of personalized medicine which aims to give ‘the right treatment to the right patient, at the right dose at the right time’ [1]. The purpose of this thesis was to acquire a deeper understanding of biomarker-guided clinical trials, by exploring and describing the various trial designs proposed to date, providing guidance on their application and on choosing the most appropriate design in a given setting, as well as considering some of the practical issues that need to be considered when implementing such trials. The finding of this thesis will help guide investigators planning biomarker-guided trials and facilitate the process of translating biomarker-discoveries into clinical practice.

The thesis begins with a comprehensive review of the literature which provided the research community with a detailed overview of the biomarker-guided clinical trial designs proposed in recent years. These designs are classified into two main categories, the so-called adaptive and non-adaptive trials. The first broad category includes eight distinct biomarker-guided adaptive designs, namely: (i) Adaptive signature design, (ii) Outcome-based adaptive randomization design, (iii) Adaptive threshold sample-enrichment design, (iv) Adaptive patient enrichment design, (v)

Adaptive parallel Simon two-stage design, (vi) Multi-arm multi-stage designs, (vii) Stratified adaptive design and (viii) Tandem two-stage design and their nine variations which were presented in detail in Chapter 2. The second broad category is composed of five main biomarker-guided non-adaptive trial designs namely: (i) single-arm designs; (ii) enrichment designs; (iii) randomize-all designs; (iv) biomarker-strategy designs and (v) other designs as well as several subtypes and extensions which were thoroughly discussed in Chapter 3. Both chapters provided clear graphical representations of each trial design, which were standardized to facilitate comparison of key features across designs, and key aspects, such as their definition, methodology, utility, advantages and disadvantages were discussed. The two chapters will serve as guidance documents for investigators embarking on biomarker-guided clinical trials and will help with addressing and reducing the confusion and ambiguity surrounding biomarker-guided designs. Key information and graphical representations of each design can also be accessed through our interactive web-tool BiGTed ([www.BiGTed.org](http://www.BiGTed.org)). It was evident that one size did not fit all and each design had a variety of both advantages and disadvantages; hence, careful consideration is needed before their implementation. This interactive web resource will improve understanding of biomarker-guided clinical trials, and will help guide researchers embarking on trials in personalized medicine in identifying the most optimal design in a given setting. The key features and user interface of BiGTed were described in detail in Chapter 4.

In Chapter 5, we explored statistical aspects of a well-known non-adaptive design, the so-called Parallel subgroup-specific design which evaluates separately the treatment effect in each biomarker-defined subgroup at the same time. We proposed sample size calculation techniques for a time-to-event outcome. Since more than one hypothesis for the assessment of efficacy of experimental treatment is being tested, the Bonferroni correction method was applied which allocates the overall level of significance between the test for biomarker-negative and biomarker-positive patients. Additionally, a simulation study was conducted with the aim of confirming



that the desirable power is achieved in each biomarker-defined subgroup under different settings of accrual and follow-up of patients. Further, the general efficiency of the study related to the cost and time of the trial was explored by implementing an adaptive approach of the aforementioned design. An interim analysis was suggested after a pre-specified percentage of events have been reached, based on which decisions are made about whether to stop the trial early due to efficacy or futility. The key issues that arose with the involvement of an interim analysis, which splits the study design into two stages, included the control of the type I error rate, pre-specification of stopping rules and stopping boundaries. Our work showed that if we use 25% of the required total number of events, the trial will be underpowered, whereas if the percentage is set at 50% and 75%, we can achieve power greater than 70% when the allocation of the level of significance is equal for both biomarker-defined subgroups. Although this chapter is based on particular settings and methods, our simulation studies provide a general insight into the operating characteristics of a design which involves an interim analysis based on a pre-specified number of events.

Chapter 6 illustrated several key aspects of biomarker-guided trial settings in practice through application to the STRONG trial, a clinical trial previously proposed to a UK funder, aiming to test whether a genotype-guided treatment strategy leads to reduced rate of relapse in alcohol-dependent patients. The Biomarker-strategy design which was initially chosen for that study, proved to be inappropriate due to the large number of patients needed to ensure sufficient power. Hence, we addressed several clinical and statistical questions to identify the most suitable non-adaptive clinical trial design for our purpose. By carefully considering the key information of each non-adaptive design described in Chapter 3 and our calculations for required total sample size for both binary and survival outcomes, we concluded that the Reverse Marker-Based strategy design is the most optimal design for the STRONG trial. In this design patients are randomized to either the genotype-guided arm where biomarker-positive patients are treated with naltrexone whereas biomarker-negative

patients are assigned to the acamprosate or to the reverse-genotype-guided arm where patients are assigned to the opposite treatment order. The aforementioned design is ethical due to the fact that both naltrexone and acamprosate have been proven effective for all patients. To complement this chapter, we incorporated the sample size re-estimation method based on unblinded interim estimates of the effect size into our selected design due to uncertainty about the true effect size in both the strategy arms. The operating characteristics of this design were investigated in Chapter 7 with the performance of two simulation studies which take into account the option of early stopping of the trial only for efficacy as well as either for efficacy or futility. Chapter 7 also provided a general overview of sample size re-estimation methods. Our findings indicated that greater power can be achieved compared to the nominal level of power when we allow the trial to stop either for efficacy or futility. However, each clinical study is different and has its own set of characteristics, therefore the conduct of simulation studies plays a key role in the understanding of the operating characteristics of a trial which will result in the right choice of trial design. The simulation codes from the R statistical software are presented in Appendix C.2, D.3.

In the final chapter of this thesis, we described various practical challenges of biomarker-guided trials, such as funding, ethical and regulatory issues, recruitment, monitoring samples and laboratories, biomarker assessment, data sharing, and resource that should be addressed when investigators conduct a biomarker-guided clinical trial. With the rapid growth of studies relating to personalized medicine, and the conduct of increasingly complex clinical trials, it is essential that these challenges are addressed, and by reflecting on existing ongoing biomarker-guided trials, we provided suggestions as a list of recommendations on how they may be resolved. Importantly, we found that collaboration between stakeholders of different backgrounds such as regulators, clinicians, statisticians etc. is essential, and could be achieved by workshops held aiming to share experiences and exchange ideas.

The current thesis enhanced the understanding of biomarker-guided clinical trials and provided much-needed guidance for stakeholders involved in the implementation of such trials. The right choice of trial design will lead to successful and efficient clinical trials in the era of personalized medicine as well as improving the drug development process which will result in safer and more effective treatments.

## 9.1. Future directions

---

In the current era of personalized medicine, biomarkers are becoming increasingly important in treatment decision-making to improve outcome for patients. Biomarkers can be binary, categorical or measured on a continuous scale. Binary biomarkers can define two subpopulations, thus, patients are classified into two biomarker-defined subgroups, the biomarker-positive and biomarker-negative subgroup, whereas categorical and continuous biomarkers can define several subpopulations. The classification of patients in the case of a continuous biomarker is more complicated since the optimal number of cutpoints and their values need to be determined. Little guidance on how to choose an optimal threshold exists in the literature and to the best of our knowledge, only one trial design specifically aimed at continuous biomarkers has been proposed in the literature, the so-called Adaptive-threshold design. This design was suggested for settings in which a putative biomarker is measured on a continuous or graded scale with its threshold for detecting individuals who would benefit from the novel treatment unknown at the initial stage of a Phase III trial (see Chapter 3). It is recommended that future research is undertaken for the investigation of methodologies to determine appropriate biomarker thresholds which will lead to optimal stratification of patients [6].

The proposed future work relating to BiGTED include incorporating sample size and power calculators for each design as well as an interactive element in which the characteristics of a study could be inserted and suggestions about the most

appropriate trial design for a given setting output. Specifically, researchers could input details about their trial e.g. Is there already strong evidence for a stratifying biomarker? YES/NO; Is the biomarker threshold already known? YES/NO etc. and the website would then provide a suggested optimal trial design.

More work is required in terms of sample size and power calculations in biomarker-guided designs. The work on Chapter 5 could be improved upon by incorporating a sample size re-estimation approach at the interim analysis stage with the aim of revising the information which is collected in case of uncertainty about the treatment effect size for which the study was powered. However, the choice of an unblinded sample size re-estimation method should be carefully evaluated as operational challenges, organizational and statistical issues may arise [3-5].

## 9.2. References

---

1. Landeck L, Kneip C, Reischl J, Asadullah K. Biomarkers and personalized medicine: current status and further perspectives with special focus on dermatology. *Experimental Dermatology*. 2016; 25(5):333-9. doi: 10.1111/exd.12948.
2. Bailey AM, Mao Y, Zeng J, Holla V, Johnson A, Brusco L, et al. Implementation of Biomarker-Driven Cancer Therapy: Existing Tools and Remaining Gaps. *Discovery medicine*. 2014; 17(92):101-14. PubMed PMID: PMC4160907.
3. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development--an Executive Summary of the PhRMA Working Group. *Journal of biopharmaceutical statistics*. 2006; 16(3):275-83; discussion 85-91, 93-8, 311-2. doi: 10.1080/10543400600614742.
4. Gallo P, Anderson K, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, et al. Viewpoints on the FDA Draft Adaptive Designs Guidance from the PhRMA Working

Group. *Journal of Biopharmaceutical Statistics*. 2010; 20(6):1115-24. doi: 10.1080/10543406.2010.514452.

5. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*. 2011; 30(28):3267-84. doi: 10.1002/sim.4102.

6. Jorgensen AL, Wason J, Kolamunnage-Dona R. Improving the efficiency of biomarker-guided trial designs by using continuous biomarker information. Grant funded by the MRC (Medical Research Council) network of hubs for trials methodology research.

## Appendix A

---

Appendix A includes supporting information related to Chapter 2 of this thesis. More specifically, variations of biomarker-guided adaptive trial designs are presented in A.1. Literature review search strategies for both biomarker-guided clinical trial designs and for traditional trial designs are given in A.2.

### A.1. Variations of biomarker-guided adaptive trial designs

---

Variations of the main biomarker-guided adaptive designs are discussed below and their characteristics are summarized further in Table A.1.

#### Variations of Adaptive signature design

##### **Adaptive threshold design**

This variation of the aforementioned Adaptive Signature design is mentioned in 25 papers (23.4%) of our review. The difference between the Adaptive Signature design and the Adaptive Threshold design is that the first one is used to develop and validate a biomarker, whereas this variant tries to identify and validate an optimal cut-off point for a pre-specified biomarker. In other words, the Adaptive Threshold design was suggested for settings in which a putative biomarker is measured on a continuous or graded scale with its threshold for detecting individuals who would benefit from the novel treatment not predefined at the initial stage of a Phase III trial. In terms of Figure 2.2, the difference between the main design (Adaptive Signature design) and this variant corresponds to the biomarker-positive subset. More precisely, in the main design, if there is no claim of treatment effectiveness in the entire population, then a portion of individuals is used to develop a predictive biomarker signature and the remaining portion is used to compare the treatment effect. However, in this variant if there is no claim of treatment effectiveness in the

entire population, the design identifies and validates a cut-off point for a prospectively selected biomarker. Adaptations here are referred to the subgroup and there are no modifications regarding the required number of patients or randomization ratio. In this design, human samples are collected to measure a pre-specified biomarker from the entire population at the beginning of the study but the value of biomarker is not used as an eligibility criteria.

Two analysis plans compose this approach, the so-called 'analysis plan A' and 'analysis plan B'. The first plan is identical to the strategy proposed for the Adaptive Signature design. The second plan uses a more effective method to accommodate the multiplicity issue when combining the statistical tests for the entire population and the biomarker-defined subgroup by incorporating the correlation structure of the two test statistics. More precisely, the plans A and B are different in the way the test statistics and thus its distribution assuming true null hypothesis is calculated. For Plan A, the test statistic is calculated as the maximum, across all possible cutoff values, of the log-likelihood ratio statistic for treatment effect for those with biomarker values above the cutoff value. For plan B, the test statistic is calculated as the larger of the test statistic for procedure A and the log-likelihood ratio statistic for treatment effect in the entire population. The second plan is considered a generalization of the first plan because in case that a difference between experimental treatment and the standard of care is demonstrated, then the next stage is to find the biomarker threshold above which the targeted treatment is more beneficial for patients than the control treatment, consequently, among the cut-off points of a measured score of a biomarker, the maximum value is selected. This plan uses a larger sample size, resulting in appropriate power for establishing the statistical significance of treatment effect restricted to patients with biomarker values above an initially unknown cut-off point.

Jiang et al. (2007) [125] proposed the use of bootstrap re-sampling method for the estimation of the point estimate and a confidence interval for the cut-off point and

described the sample size planning for this design. Also, according to Jiang et al. (2007) [125] if plan B does not reject the null hypothesis, the estimation of the cut-off point value would be inexplicable, and thus it should not be estimated. Jiang et al. (2007) [125] demonstrated through a simulation study that the second plan was more effective than the first one. A modification of the global test of the null hypothesis which was used by Jiang et al. (2007) [125] is illustrated in the paper of Simon (2012) [57]. The Adaptive Threshold design can detect efficiently a global treatment effect and provides statistically valid tests when the promising treatment effect is limited to a particular biomarker-defined subset, however, a larger sample size may be required and can lead also to redundant power.

### **Molecular signature design**

Molecular Signature design is mentioned in 2 two articles (1.9%). It is a Phase III design which collects tissue samples from the entire population at the start of the trial and analyse them when the study is near completion.

After the collection of tissue samples from the entire population, all patients are randomized to either the experimental treatment or the standard treatment. The methodology is similar to the Adaptive Signature design. This approach makes the comparison of the novel drug with the standard of care, but on a primary outcome measure which here is the overall survival using the significance level of 0.04. In case that the results show the effectiveness of an experimental treatment over the control arm, we claim the effectiveness of treatment in the overall population. Otherwise, an analysis is conducted for the identification and validation of the biomarker classifier (i.e. a combination of biomarkers) which gives the best primary outcome measure. A portion of subjects is used for the detection of a biomarker classifier and the remainder of patients for its validation. It is considered as a promising strategy without statistical considerations mentioned.



### **Cross-validated Adaptive Signature design**

Cross-validated Adaptive Signature design (CVASD) is found in 19 papers (17.8%) of our review. It was proposed by Freidlin et al. (2010) [60] aiming to increase the efficiency of the Adaptive Signature design. Similar to the Adaptive signature approach it is a Phase III frequentist trial design based on a fall back strategy in order to identify candidate biomarkers in the training set of the study and evaluate them in the validation set.

The difference between Adaptive signature design and Cross-validated Adaptive Signature design is in terms of the methodology analysis. The former is composed of a split-sample approach, using approximately half of patients to develop the biomarker signature and the remainder of patients to validate it, whereas, the latter uses the K-fold cross validation procedure, i.e. there are K cross-validated training sets which are used to classify subjects in the corresponding K cross-validated validation sets. After the classification of all patients, we compare the experimental treatment versus the control treatment in the biomarker-positive patients (i.e. subgroup of classifier positive patients). The Cross-validated Adaptive Signature design may yield larger power but it faces the same challenges with its main design and also includes the multiplicity problem.

### **Generalized adaptive signature design**

Generalized Adaptive Signature approach is described in 2 papers (1.9%). Firstly, candidate biomarkers are selected and the cut-off points are optimized using a training set and secondly, the chosen biomarkers are assessed in the validation set. According to Simon (2010) [18], this approach is applicable when there are a number of available candidate biomarkers, but data from a Phase III setting is required for choosing the most appropriate biomarkers. A major drawback of this design is the limited power when we assess the treatment effect in the biomarker-defined subset.

### **Adaptive signature design with subgroup plots**

Adaptive Signature design with Subgroup Plots [64] is an extension of Adaptive Signature design which has been proposed in order to add flexibility. It uses tail-oriented or sliding window subgroup plots in order to identify a subset of patients which is most likely to respond to a particular experimental treatment after taking into account several cut-off points of the benefit score obtained by the subgroup plots. In this way it provides broader confidence intervals of the estimated treatment benefit. No statistical considerations have been found for this approach.

### **Variation of Outcome-based adaptive randomization design**

#### **Bayesian covariate adjusted response-adaptive randomization**

The Bayesian Covariate Adjusted Response-Adaptive Randomization (BCARA) is identified in two articles (1.9%) and it was proposed in 2010 by Eickhoff et al. (2010) [53]. This strategy which combines a Bayesian, an adaptive and biomarker classification approach aims to match patients with the most efficacious treatments by utilizing patient's biomarker information becoming available during the conduct of the clinical trial. This strategy may be useful in the explanatory phase II setting of the drug development [53]. It is also considered as a response-adaptive randomization strategy as the allocation of the study population depends on the responses of previous outcomes. A partial least square logistic regression approach is conducted to determine adaptively predictive biomarker-defined subsets.

The general procedure of this approach is composed of four steps according to Eickhoff et al. (2010) [53]: (i) randomly assign the first  $n^* \geq J^*(K + 1)$  patients to the different treatment arms where  $J$  the number of different treatment groups and  $K$  the number of biomarkers. At least one response should be observed in each of the different treatment groups before moving to the Bayesian response adaptive randomization; (ii) after each new individual has been enrolled in the study,

predictive biomarker-defined groups are determined by utilizing a partial least squares logistic regression strategy (PLSLR) which can predict whether the patient can benefit from the treatment. The biomarker status is determined before the randomization; (iii) after the establishment of the biomarker status and biomarker-defined groups of each new individual, the individual is then randomly assigned into one of the treatment arms using a BCARA randomization; (iv) according to the results of the BCARA randomization the trial either stops or continues based on decision rules proposed by Eickhoff et al. (2010) [53]. The Bayesian covariate adjusted response-adaptive trial design has the ability to identify the biomarker-defined groups likely to respond to a treatment but it does not control the Type I error and in order to ensure that the identified result is true, a Phase III study should be conducted.

#### **Variation of Adaptive patient enrichment design**

##### **Modified Bayesian version of the two-stage design of Wang et al. (2007) [80]**

A variation of Adaptive Patient Enrichment design by Wang et al. (2007) [80] was found in 2 papers (1.9 %). It is a Phase III Bayesian two-stage design proposed by Karuri and Simon (2012) [7] for the evaluation of both treatment and biomarker.

Karuri and Simon (2007) [7] use a Bayesian framework in order to allow further flexibility for expressing the degree of prior information regarding the utility of a biomarker. More precisely, posterior distribution of treatment effects within the biomarker-positive and biomarker-negative subgroups based on an interim analysis of first-stage is used in order to come to a decision regarding the recruitment and the continuation of the trial. This approach allows for early termination of the study during the initial stage of the trial and has a satisfactory power. No statistical challenges have been identified.

#### **Variations of Multi-arm multi-stage (MAMS) design**

## Two-stage adaptive seamless design

Two-stage Adaptive Seamless design is a type of clinical trial design identified in 28 papers (26.2%) of our review. It uses the MAMS approach combining two separate studies into one single study and uses interim monitoring as well as multi-arm design features. It connects the explanatory Phase II stage for treatment selection and confirmatory Phase III stage for the final comparison of the chosen experimental treatments with the standard of care. The Two-stage Adaptive Seamless design aims to improve the power in the Phase II stage in order to continue on the Phase III stage having obtained important promising information. In the definitive analysis, it uses data from patients registered during the Phase II and Phase III stages. A prerequisite of this strategy is the availability of a reliable early endpoint. An example of actual trial which uses the two-stage adaptive seamless design is the ISPY2 trial [54, 93, 126, 127].

Brannath et al. (2009) [36] propose an approach which uses Bayesian decision tools and is based on the two-stage seamless design in order to confirm that the identified biomarker-defined subgroup from a Phase II study is sensitive to the new treatment in a separate explanatory phase (i.e. a study which is conducted at the same time with the two-stage adaptive seamless design) and afterward conduct a Phase III study with this selected subgroup. More precisely, the general procedure of this Phase II/III strategy is presented by Brannath et al. (2009) [36] as follows: When half of individuals are recruited in the study, an interim analysis is performed in order to decide whether to accept or not a biomarker-defined subpopulation identified in a separate exploratory study. At this interim stage, a decision is also made about whether to continue accruing patients from the aforementioned biomarker-defined subset or from the entire study population. If the first case occurs, the treatment effect is assessed only in this biomarker subpopulation and if the second case happens, the treatment effect is tested in the entire population and biomarker-defined subgroup at the same time. In case that there is no identified biomarker-defined subpopulation

from the separate exploratory study, the trial continues in the overall population using a classical group sequential design. The major advantage of this type of design is its ability to reduce the costs and also the selection of the target population in a reliable way. Also, appropriate methodology, such as that used by Brannath et al. (2009) [36] where multiple testing is adjusted by a weighted combination of p-value from data of the second stage and the first stage, and Simes' step-up process [128] is used when combining data from both Phases in order to maintain the Type I error rate. An extension of the above approach by Brannath et al. (2009) [36] is proposed by Jenkins et al. (2011) [129] which can result in the rapid approval of novel treatments to the most appropriate individuals who are likely to benefit from the new drug. During the Phase II trial an interim analysis is conducted using a short-term intermediate outcome measure (i.e. survival endpoint) in order to select the population (either the entire population or the biomarker-positive patients) which will be used in the Phase III study with a long-term endpoint.

Mehta et al. (2014) [130] proposed an alternative seamless approach for subgroup selection in time-to-event-data for situations where there is no a priori assumption that a biomarker is predictive of treatment efficacy; consequently their design tests whether there is treatment effect in both biomarker-negative and biomarker-positive subpopulation separately instead of testing the null hypothesis of no treatment effect in the entire study population and in biomarker-positive subset.

According to Scher et al. (2011) [59], formulas for sample size calculation/allocation are proposed in situations where the study endpoints are continuous, discrete, and contain time-to-event data supposing the availability of a well-established relationship between the study endpoints at different stages, and that the study objectives at different stages are the same. Ang et al. (2010) [52] have stated that even in case that the trial stops early, a Phase III infrastructure should be developed. Such strategies have been proposed by Ellenberg and Eisenberger (1985) [131] and Inoue et al. (2002) [132] for evaluating the possibility to stop early or to

continue to the confirmatory phase III repeatedly during the explanatory phase. The aforementioned designs are useful in situations where there is strong belief in the efficacy of the experimental therapy that can lead the study to the confirmatory phase, but confirm of this assumption is needed [52]. Despite the fact that this approach is considered as a more efficient strategy yielding larger power as compared with the conduct of separate trials, it can lead to introduction of bias and inflation of the type I error rate.

### **Group Sequential design**

Adaptive Group Sequential design is found in 2 papers (1.9%) which can be incorporated into the MAMS approach for the development and validation of personalized therapies and is proposed by Lai et al. (2013) [74]. This strategy aims to find the most beneficial treatment for future patients based on their biomarker profiles, with a guaranteed probability of correct selection. It was proposed for the examination of multiple composite hypotheses not only in the entire study population but also in the biomarker-positive subgroups [133]. The design is based on approved treatments, and aims to improve patient's health by providing them with the most efficacious (yet unidentified) treatment. Additionally, another crucial objective of this approach is that the development of a novel treatment strategy for the forthcoming patients and the confirmation that the treatment effect of this strategy is in fact more effective than the historical mean effect of the control treatment plus a predetermined threshold [74].

According to Lai et al. (2013) [74], it is an approach for “jointly developing and testing treatment recommendations for biomarker classes, while using multi-armed bandit ideas to provide sequentially optimizing treatments to patients in the trial”. According to an interim data analysis, sequential decisions about whether to continue the study or not, are taken. It is considered a simple approach where selection of cut-off points is not required before the conduct of the first interim analysis.

**Table A.1.** Characteristics of variations of Biomarker-guided adaptive trial designs

Types of variations of Biomarker-guided adaptive trial designs	Phase	Pros	Cons
<p><b><u>Adaptive threshold design (25 papers)</u></b> [3, 6, 8, 12, 14, 15, 18, 20, 21, 27, 29, 30, 47, 57, 58, 63, 64, 68, 70, 74, 78, 84, 125, 134]</p> <p><b>Also called:</b></p> <p>Biomarker adaptive threshold design</p>	III	<p>Validation of a candidate biomarker without need for an established cut-off point.</p> <p>Identification of an optimal cut-off point for detecting sensitive patients (i.e. biomarker-positive patients).</p> <p>Detection of overall treatment effect if one exists.</p> <p>Statistically valid test if treatment benefit is restricted to a biomarker-defined subgroup.</p> <p>Reduces dependence on Phase II data for establishing a test cut-off point.</p> <p>More efficient design as compared to the traditional design (i.e. standard broad eligibility Phase III design based on assessing the global treatment effect in the</p>	<p>Requirement of a pre-specified biomarker for sensitivity, but not an established cut-off point.</p> <p>Data from the same study to both define and validate the cut-off point of the biomarker may raise concerns.</p> <p>Augmented costs due to the potential sample size increase and/or redundant power by partitioning the overall type I error.</p>

overall population when the proportion of sensitive patients is low).			
<u><b>Molecular signature design (2 papers)</b></u> [32, 63]  No alternative names found for this trial design	III	Considered as a promising strategy for drug development as it takes advantage of the use of an end-point with clear clinical gain.	No information found
<u><b>Cross-validated adaptive signature design (19 papers)</b></u> [9, 14-16, 18, 20, 21, 24, 27, 32, 59, 60, 62-64, 66, 84, 104, 135]  No alternative names found for this trial design	III	Gain more power as it could maximize the number of individuals taking part in the development of the biomarker signature.  Can detect the subset of patients most likely to respond to a specific treatment in a more reliable way.	Same challenges as the Adaptive signature design.  Multiplicity problem for statistical testing as the statistical test would be conducted twice.



<b><u>Generalized adaptive signature</u></b> (2 papers) [18, 63]  No alternative names found for this trial design	III	Optimizes the test based on randomized data for patients in the Phase III setting.	Limits its power when testing the effectiveness of an experimental treatment in the biomarker-positive subgroup.
<b><u>Adaptive signature design with subgroup plots</u></b> (1 paper) [64]  No alternative names found for this trial design	III	No information found	No information found
<b><u>Bayesian covariate adjusted response-adaptive randomization</u></b> (2 papers) [53, 63]	II	Ability to incorporate prior knowledge from biomarkers into the design.  Identification of the subgroups for which a particular experimental treatment is more effective.	The Type I error is not controlled in the traditional sense.  An independent Phase III study focused on the selected biomarker-defined subgroups is required to show that the identified promising result is definitely true.

<p>No alternative names found for this trial design</p>	<p>Can result in reduction of the number of patients required when compared to alternative designs (i.e. non-adaptive trial designs).</p> <p>Solves the issue of the incorporation of information of multiple and possibly correlated biomarkers.</p>	
<p><b><u>Modified Bayesian version of the two-stage design of Wang et al. (2007) [80] (2 papers) [7, 136]</u></b></p> <p><b>Also called:</b></p> <p>Two-Stage Bayesian design</p>	<p>III</p> <p>Can incorporate prior belief regarding the strength of biomarker into the Phase III setting using a Bayesian framework and simultaneously protecting the study population and minimizing the Type I error in the biomarker-positive and biomarker-negative subgroups</p> <p>Can terminate the study early according to whether the treatment is effective or not in the biomarker-positive subgroup at the interim stage whereas the main design by Wang et al. (2007) [80] does not allow for early termination of the trial.</p> <p>Satisfactory power for testing the biomarker-positive subgroup.</p>	<p>No information found</p>

		<p>Enables the reduction of number of biomarker-negative patients for whom a particular treatment tailored to them seems to be ineffective according to biological evidence.</p> <p>The utilized Bayesian formulation sheds light on the nature of inference at the end of the study.</p> <p>Can result in reduction of costs of clinical development.</p>	
<p><b><u>Two-stage adaptive seamless design</u></b> (28 papers) [20, 23, 33, 36, 37, 40, 42, 43, 45-47, 52, 54, 59, 68, 74, 82, 93, 119, 126, 129-131, 137-142]</p> <p><b>Also called:</b></p> <p>Seamless Phase II/III designs</p> <p>Adaptive Seamless</p>	II/III	<p>The evaluation of each experimental therapy can be performed without requiring the conduct of separate large-scale Phase II trials.</p> <p>Flexibility and efficiency of trials can be increased.</p> <p>Individuals from both explanatory and confirmatory stages are used in the definitive analysis; hence, the design avoids ‘wasting’ individuals already registered in Phase II setting.</p>	<p>A significant concern is that the results obtained from the Phase II analysis, which becomes an interim analysis of the Phase III study should remain in the hands of the data monitoring committee due to confidentiality.</p> <p>Concerns also arise regarding the efficiency and validity of such a trial design.</p> <p>Sometimes, the endpoints within the different phases are dissimilar, hence, a decision is required on how to combine the data obtained from both stages in order to use them in the definitive analysis. This challenge is further discussed in the paper of Chow et al. (2007) [141].</p>

Phase II/III Adaptive design	Diminishes the potential loss of time between the completion of Phase II stage and the beginning of patient enrollment in Phase III setting.	According to the Draft Guidance by U.S. Food and Drug Administration (2010) [142], an adaptive seamless phase II/III design is described as a less well-understood design which may introduce bias and inflation of the Type I error rate.
Two-stage Adaptive Seamless design	The same standard of care can be used in both stages of the study.	
Adaptive Seamless Phase II/III design	<p>Can result in the same quality of evidence as in a traditional design but with a smaller number of patients.</p> <p>Can result in the speedup of drug development and also in a successful Phase III trial.</p> <p>More efficient due to the improved power and the ability to control the Type I error as compared with the conduct of separate studies.</p> <p>The required power of both individual studies may be acquired by using a smaller number of patients than that of a single study.</p> <p>Two separate trials (Phase II and Phase III) are conducted under one single trial protocol, but in reality, researchers</p>	Calculation and allocation of the necessary sample size for the two separate studies.

	analyze them separately using data from each stage, resulting in this way in savings of time and cost.	
<b><u>Group Sequential design</u></b> (2 papers) [74, 133]	Researchers do not have to choose the cut-off points which should be used to designate the biomarker classes until the performance of the first interim analysis.	No information found
No alternative names found for this trial design	Non-promising treatments can be dropped at an early stage.	

## A.2. Literature review search strategies for both biomarker-guided clinical trial designs and for traditional trial designs

---

### *MEDLINE*

#### **Traditional clinical trial designs**

1. Clinical Trials as Topic/
2. Clinical Trial/
3. 1 or 2
4. Research Design
5. design\*.mp.
6. Research design\*.mp.
7. Statistical design\*.mp.
8. Study design\*.mp.
9. Traditional design\*.mp.
10. Trial design\*.mp.
11. 4 or 5 or 6 or 7 or 8 or 9 or 10
12. 3 and 11
13. limit 12 to (english language and "review articles")
14. limit 13 to last ten years

## ***MEDLINE***

### **Biomarker-guided clinical trial designs**

1. Clinical Trials as Topic/
2. clinical trial\*. ti, ab.
3. 1 or 2
4. design\*. ti, ab.
5. 3 and 4
6. limit 5 to comment
7. limit 5 to editorial
8. limit 5 to journal article
9. limit 5 to guideline
10. limit 5 to systematic reviews
11. limit 5 to "review"
12. limit 5 to technical report
13. limit 5 to practice guideline
14. 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13
15. limit 14 to English language
16. limit 15 to last 10 years
17. exp \*marker/ or biological marker/ or clinical marker/

18. (marker\* or biomarker\* or factor\* or classifier or signature\* or target\* or endpoint). ti, ab.

19. 17 or 18

20. 16 and 19

The Ovid strategy was conducted by following the guidance by BMA Library - MEDLINE Plus. Basic Course. Notes for OvidSP; 2012. Available from: [http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0ahUKEwjS7\\_OmodvJAhWGVhQKHZr0AZMQFggdMAA&url=http%3A%2F%2Fbma.org.uk%2F-%2Fmedia%2Ffiles%2Fpdfs%2Fabout%2520the%2520bma%2Flibrary%2Fmedline%2520plus%2520basic%2520course%2520manual%25202012.pdf&usg=AFQjCNGFxcWiS11CJsroeeIETAWjW0neUA](http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0ahUKEwjS7_OmodvJAhWGVhQKHZr0AZMQFggdMAA&url=http%3A%2F%2Fbma.org.uk%2F-%2Fmedia%2Ffiles%2Fpdfs%2Fabout%2520the%2520bma%2Flibrary%2Fmedline%2520plus%2520basic%2520course%2520manual%25202012.pdf&usg=AFQjCNGFxcWiS11CJsroeeIETAWjW0neUA).



## Appendix B

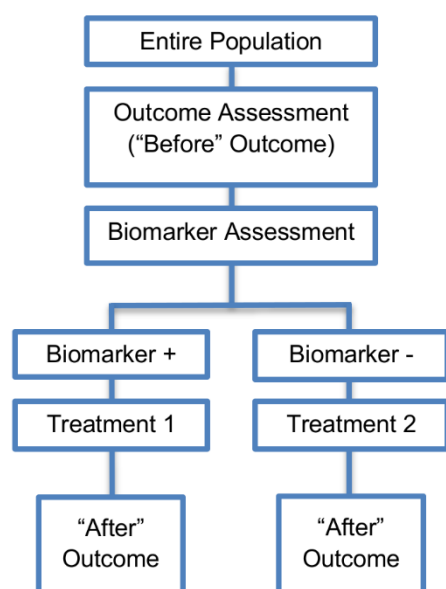
---

Appendix B includes supporting information related to Chapter 3. More precisely, extensions of biomarker-guided non-adaptive trial designs are given in the current section. The literature review search strategies for both biomarker-guided clinical trial designs and for traditional trial designs are given in A.2.

Extensions of biomarker-guided non-adaptive trial designs:

### *Variations of Biomarker-Strategy Designs*

**Sequential before-after pharmacogenetic diagnostic study:** The design identified in two papers [36,38] (2%) of our review. A graphical illustration of this design is given in Figure B.1.

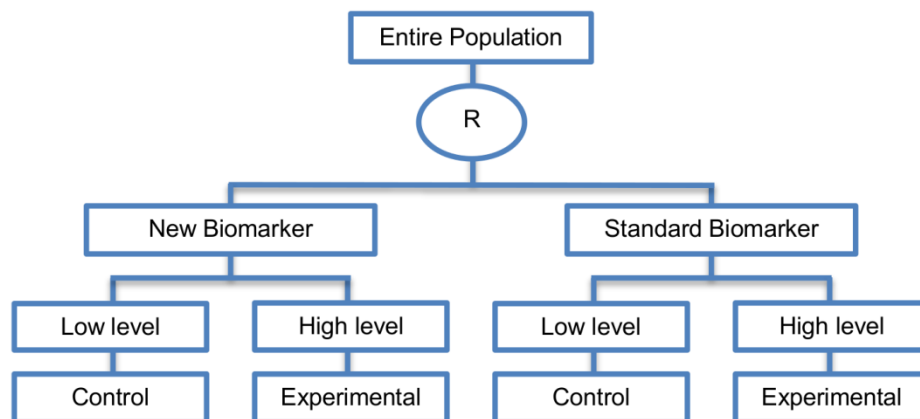


**Figure B.1.** Sequential before–after pharmacogenetic diagnostic study

This approach was proposed in the field of pharmacogenetics involving the assessment of pharmacogenetic diagnostics being performed during the study. In this sequential approach each patient serves as his/her own control) [38]. Treatments are tailored to patients before genotyping and then again after genotyping. A comparison of outcomes before and after the introduction of pharmacogenomics is conducted.

This individual crossover approach requires a smaller number of patients as compared to the previous designs described and is not considered complex in its implementation. Additionally, before–after comparisons are the basis of medical practice in many important therapeutic areas such as surgery [38] and they can inform researchers about whether a personalized treatment is more effective than the standard of care. However, types of systematic error (i.e. bias which yield incorrect estimate of a measure of disease) might be introduced, e.g., due to errors on classification of outcomes or on the assessment of the biomarker status of patients.

**Classifier randomization design:** Another extension of biomarker-strategy designs is the Classifier randomization design which was identified in two papers [36,107] (2%) of our review. An example of an actual trial which uses this strategy is the NNBC-3 European trial [107]. A graphical illustration of this design is given in Figure B.2.

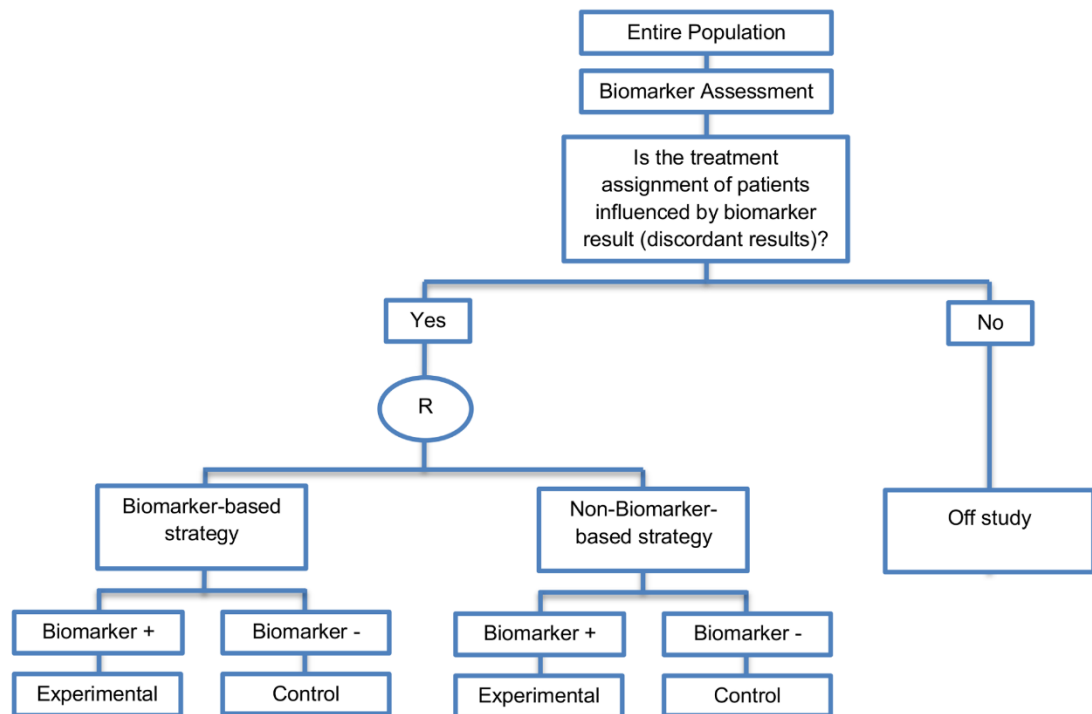


**Figure B.2.** Classifier randomization design. “R” refers to randomization of patients.

This is an approach proposed for the validation of prognostic biomarkers which randomly assign the study population between classifiers (i.e. a new biomarker and a standard biomarker) rather than treatments validating directly the new biomarker. With this approach, we compare the new and standard classifiers in order (i) to show equivalence in outcome, i.e. the outcome of patients assigned to the new biomarker is not too different to that obtained from patients in the standard classifier independently of the low/high level in each category; (ii) to show superiority in outcome, i.e. the outcome of patients assigned to the new classifier who are given the

experimental treatment is superior to that obtained from the patients given the control treatment.

**Modified marker strategy design:** This modified version of the biomarker-strategy design was identified in five papers [9,19,22,58,91] (5%) of our review. A graphical illustration of this approach is given in Figure B.3.



**Figure B.3.** Modified marker strategy design. “R” refers to randomization of patients.

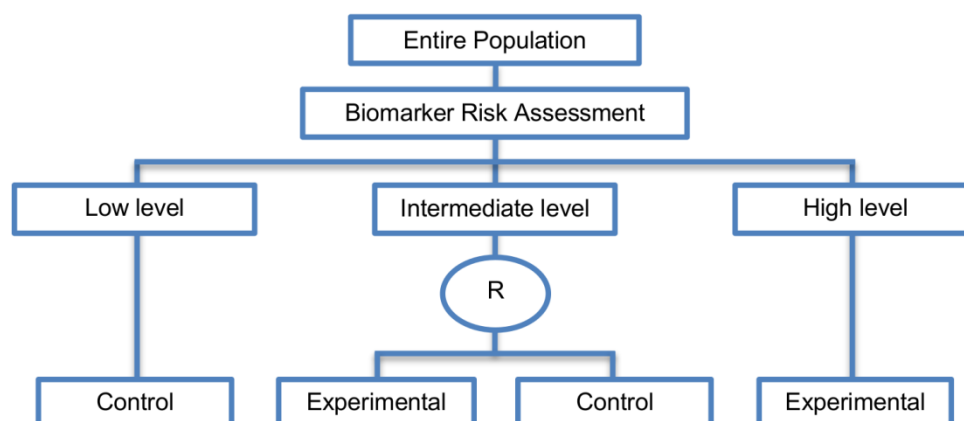
In this design, biomarker assessment is required in the entire population before randomization to either the biomarker-based strategy arm or to the non-biomarker-based strategy arm. However, only patients for whom treatment assignment would be influenced by biomarker result are then randomized—those whose treatment assignment would be the same regardless of biomarker result are off study. An example of an actual trial which uses this approach is the MINDACT trial. More precisely, in the MINDACT trial the entire population is evaluated by two analyses, the 70-gene profile (biomarker-based analysis) as well as clinicopathologic factors (non-biomarker-based analysis). Next, only patients with discordant results (i.e. patients predicted to be high risk based on one of the two analyses and low risk

patients based on the other analysis) will be randomized to either the biomarker-based strategy or to the non-biomarker-based strategy).

Next, as exactly in the biomarker-strategy design with biomarker assessment in the control arm, patients who are randomly assigned to the biomarker-based strategy arm are treated either with the experimental treatment if they have biomarker-positive status or with the control treatment if they have biomarker-negative status. The main limitation of biomarker-strategy designs is that they can result in a significant number of patients in the trial who are assigned to the same treatment in both biomarker-based strategy arm and non-biomarker-based strategy arm (i.e. biomarker-negative patients in the biomarker-based strategy arm receive control treatment but, biomarker-negative patients might also receive control treatment in the non-biomarker-based strategy arm as the random assignment of the entire population to this strategy arm is independent of their biomarker-status). Consequently, a large sample size is needed to identify the diluted treatment effect. This modified version promises to solve this limitation of biomarker-strategy designs by only randomizing patients for whom treatment assignment is influenced by biomarker result.

### *Variation of Randomize-All Designs*

**Two-way stratified design:** A version of the Biomarker Stratified design is the Two-way Stratified design which was referred to in two papers [107,132] (2%) of our review. Figure B.4 represents the graphical illustration of this strategy.



**Figure B.4.** Two-way Stratified design (for validation of prognostic biomarkers). “R” refers to randomization of patients.

In this approach, patients are stratified according to their biomarker results. Standard treatment is tailored to patients with a low score, experimental treatment is given to patients with a high score and those with an intermediate score are randomized to either experimental or standard treatment. An example of an actual trial which uses this approach is that conducted by the “Arbeitsgemeinschaft Gynakologische Onkologie” (AGO) Study Group in cooperation with the EORTC Receptor and Biomarker Study Group for the “Chemo-N0 trial” to validate UPA and PAI1 as prognostic indicators in node negative breast cancer [107].

According to Spira et al. [132] a noninferiority design for the intermediate group can be used and has the statistical power to detect a 3% or greater difference between the randomized arms.

Since there is no direct prospective comparison between the novel and standard classifier, the two-way stratified design provides further, although indirect, validation of the biomarker.

## Appendix C

---

Appendix C includes supporting information (table and figures) and R codes related to Chapter 5. Tables and figures can be found in C.1. R codes used to produce the results of the Parallel Subgroup-Specific design and its adaptive version are given in C.2.1 and C.2.2 respectively.

### C.1. Supporting tables and figures

---

**Table C.1.** Accrual rate and number of events and patients (calculated from (5.7), (5.1) and (5.3) respectively) which achieve approximate 80% power for different scenarios of hazard ratios and significance levels, and the corresponding power of each biomarker-defined subgroup yielded from the simulation.

	Simulation setting		Accrual rate and required numbers			Simulated power
Group of patients	Significance level	Hazard ratio	Accrual rate	Number of events	Number of patients	Power
Biomarker-negative	0.0125	0.6	9	146	168	0.798
Biomarker-positive	0.0125	0.4	4	45	76	0.788
Entire population	0.025	-	-	191	244	-

Biomarker-negative	0.015	0.6	9	139	160	0.796
Biomarker-positive	0.010	0.4	4	48	81	0.792
Entire population	0.025	-	-	187	241	-
Biomarker-negative	0.010	0.6	10	154	177	0.791
Biomarker-positive	0.015	0.4	4	43	73	0.792
Entire population	0.025	-	-	197	250	-
Biomarker-negative	0.0125	0.7	19	299	335	0.804
Biomarker-positive	0.0125	0.5	7	79	126	0.786
Entire population	0.025	-	-	378	461	-
Biomarker-negative	0.015	0.7	18	285	320	0.798
Biomarker-positive	0.010	0.5	7	84	133	0.793
Entire population	0.025	-	-	369	453	-

Biomarker-negative	0.010	0.7	20	316	354	0.798
Biomarker-positive	0.015	0.5	7	76	120	0.800
Entire population	0.025	-	-	392	474	-
Biomarker-negative	0.0125	0.8	47	764	841	0.796
Biomarker-positive	0.0125	0.6	12	146	220	0.796
Entire population	0.025	-	-	910	1061	-
Biomarker-negative	0.015	0.8	45	729	803	0.800
Biomarker-positive	0.010	0.6	13	154	232	0.796
Entire population	0.025	-	-	883	1035	-
Biomarker-negative	0.010	0.8	49	806	888	0.799
Biomarker-positive	0.015	0.6	12	139	210	0.793
Entire population	0.025	-	-	945	1098	-



Biomarker-negative	0.0125	0.9	207	3425	3720	0.806
Biomarker-positive	0.0125	0.7	24	299	434	0.804
Entire population	0.025	-	-	3724	4154	-
Biomarker-negative	0.015	0.9	197	3268	3550	0.803
Biomarker-positive	0.010	0.7	25	316	458	0.798
Entire population	0.025	-	-	3584	4008	-
Biomarker-negative	0.010	0.9	218	3616	3928	0.804
Biomarker-positive	0.015	0.7	23	285	414	0.796
Entire population	0.025	-	-	3901	4342	-

**Table C.2.** Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 2 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the second scenario of hazard ratios and significance levels are also presented.

		Simulation setting		Number		Simulated Power					
Group of patients	Significance level	Hazard ratio	Required	Required	Expected Total	Expected	Expected	FSP	ESP	Power	
			Number of events	Number of patients	study period (months)	Number of events	Number of patients				
25%	Biomarker-negative	0.0125	0.7	299	335	17.6	176	197	0.3672	0.1817	0.5678
	Biomarker-positive	0.0125	0.5	79	125	16.9	44	70	0.3931	0.1895	0.5382
	Entire population	0.025	-	378	460	-	220	267	-	-	-
	Biomarker-negative	0.015	0.7	285	320	16.8	160	179	0.4173	0.1696	0.5000
	Biomarker-positive	0.010	0.5	84	133	14.0	39	62	0.5070	0.2060	0.4258
	Entire population	0.025	-	369	453	-	199	241	-	-	-
	Biomarker-negative	0.010	0.7	316	354	14.9	157	176	0.4823	0.1879	0.4450

	Biomarker-positive	0.015	0.5	76	121	16.1	41	65	0.4268	0.1899	0.5028
	Entire population	0.025	-	392	475	-	198	241	-	-	-
50%	Biomarker-negative	0.0125	0.7	299	335	21.7	217	243	0.1655	0.3848	0.7246
	Biomarker-positive	0.0125	0.5	79	125	21	55	88	0.1846	0.4142	0.7004
	Entire population	0.025	-	378	460	-	272	331	-	-	-
	Biomarker-negative	0.015	0.7	285	320	21.3	203	228	0.2036	0.3736	0.6700
	Biomarker-positive	0.010	0.5	84	133	19.6	55	87	0.2674	0.4273	0.6145
	Entire population	0.025	-	369	453	-	258	315	-	-	-
	Biomarker-negative	0.010	0.7	316	354	20.2	212	238	0.2425	0.4127	0.6414
	Biomarker-positive	0.015	0.5	76	121	21.0	53	85	0.2118	0.3915	0.6593
	Entire population	0.025	-	392	475	-	265	323	-	-	-

75%	Biomarker-negative	0.0125	0.7	299	335	25.0	249	280	0.0707	0.5922	0.7739
	Biomarker-positive	0.0125	0.5	79	125	24.8	65	104	0.0824	0.6046	0.7571
	Entire population	0.025	-	378	460	-	314	384	-	-	-
	Biomarker-negative	0.015	0.7	285	320	25.0	238	267	0.0944	0.5671	0.7264
	Biomarker-positive	0.010	0.5	84	133	24.2	68	107	0.1322	0.6366	0.7193
	Entire population	0.025	-	369	453	-	306	374	-	-	-
	Biomarker-negative	0.010	0.7	316	354	24.5	258	289	0.1141	0.6187	0.7262
	Biomarker-positive	0.015	0.5	76	121	24.9	63	100	0.1007	0.5812	0.7277
	Entire population	0.025	-	392	475	-	21	389	-	-	-

**Table C.3.** Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 3 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the third scenario of hazard ratios and significance levels are also presented.

		Simulation setting		Number		Simulated Power					
25%	Group of patients	Significance level	Hazard ratio	Required Number of events	Required Number of patients	Expected Total study period (months)	Expected Number of events	Expected Number of patients	FSP	ESP	Power
	Biomarker-negative	0.0125	0.8	764	841	17.6	449	495	0.3678	0.1815	0.5680
	Biomarker-positive	0.0125	0.6	146	221	16.9	82	125	0.3921	0.1903	0.5399
	Entire population	0.025	-	912	1061	-	531	620	-	-	-
	Biomarker-negative	0.015	0.8	729	803	16.8	409	450	0.4162	0.1694	0.5004
	Biomarker-positive	0.010	0.6	154	233	14	72	109	0.5072	0.2051	0.4245
	Entire population	0.025	-	883	1036	-	481	559	-	-	-
Biomarker-negative	0.010	0.8	806	888	14.9	400	440	0.4839	0.1881	0.4434	

	Biomarker-positive	0.015	0.6	139	210	16.1	75	113	0.4276	0.1893	0.5013
	Entire population	0.025	-	945	1098	-	475	553	-	-	-
50%	Biomarker-negative	0.0125	0.8	764	841	21.7	554	609	0.1658	0.3849	0.7244
	Biomarker-positive	0.0125	0.6	146	221	21.0	102	155	0.1827	0.4161	0.7034
	Entire population	0.025	-	912	1061	-	656	764	-	-	-
	Biomarker-negative	0.015	0.8	729	803	21.3	518	571	0.2036	0.3743	0.6711
	Biomarker-positive	0.010	0.6	154	233	19.6	101	152	0.2687	0.4245	0.6118
	Entire population	0.025	-	883	1036	-	619	723	-	-	-
	Biomarker-negative	0.010	0.8	806	888	20.2	542	597	0.2438	0.4123	0.6400
	Biomarker-positive	0.015	0.6	139	210	21	97	147	0.2126	0.3895	0.6576
	Entire population	0.025	-	945	1098	-	639	744	-	-	-

75%	Biomarker-negative	0.0125	0.8	764	841	25.1	637	702	0.0706	0.5923	0.7739
	Biomarker-positive	0.0125	0.6	146	221	24.8	121	183	0.0821	0.6053	0.7580
	Entire population	0.025	-	912	1061	-	758	885	-	-	-
	Biomarker-negative	0.015	0.8	729	803	25.0	608	670	0.0944	0.568	0.7276
	Biomarker-positive	0.010	0.6	154	233	24.2	124	188	0.1345	0.6336	0.7160
	Entire population	0.025	-	883	1036	-	732	858	-	-	-
	Biomarker-negative	0.010	0.8	806	888	24.5	658	725	0.1144	0.6199	0.7252
	Biomarker-positive	0.015	0.6	139	210	24.9	115	175	0.1021	0.5755	0.7236
	Entire population	0.025	-	945	1098	-	773	900	-	-	-

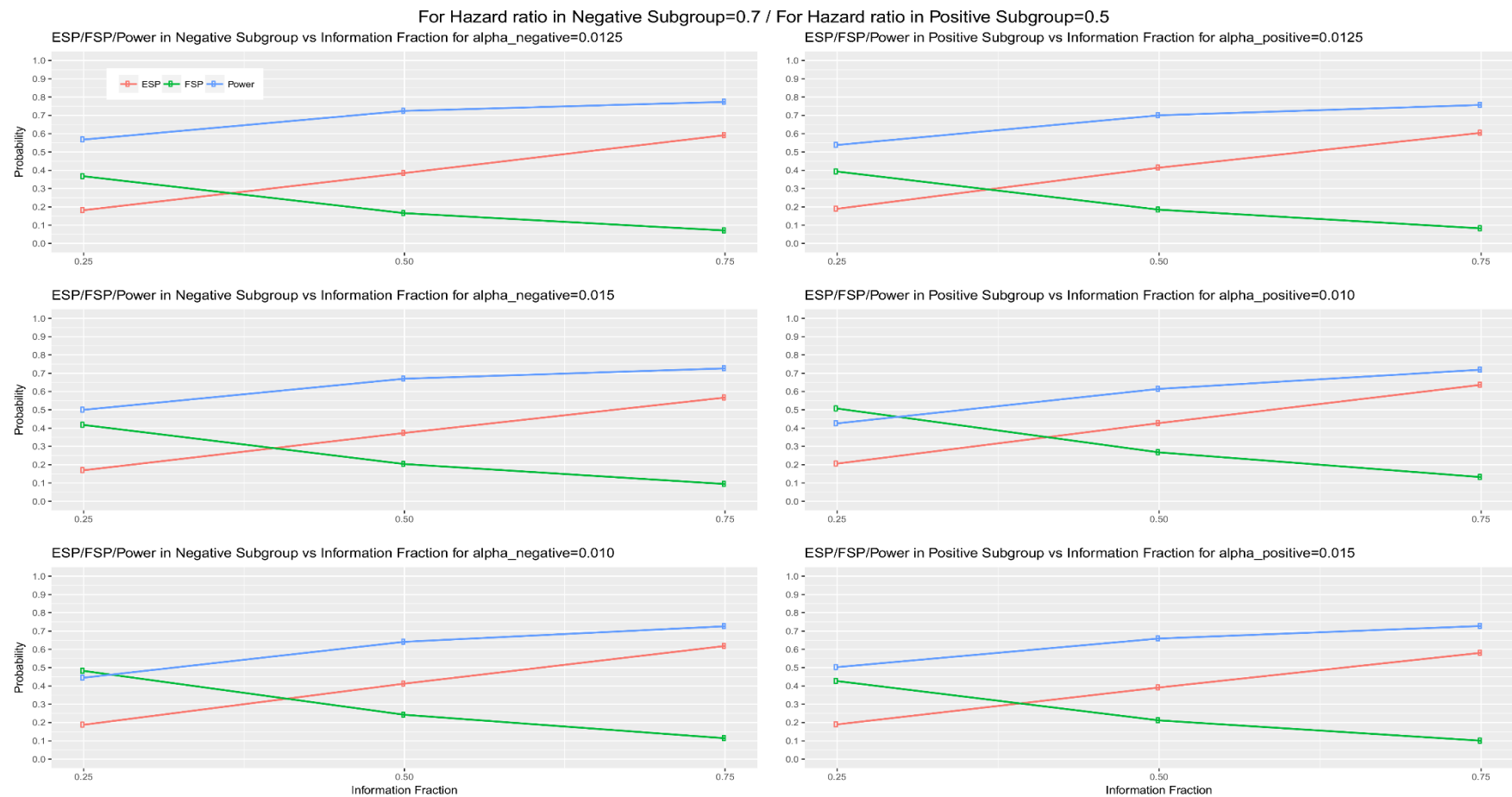
**Table C.4.** Results for expected number of events and patients, expected total study period, futility stopping probability, efficacy stopping probability and power of a two-stage design in scenario 4 of hazard ratios for different percentages of information fraction. Number of events and patients from Table C.1 (calculated from (5.1) and (5.3) respectively) which achieve 80% power for the fourth scenario of hazard ratios and significance levels are also presented.

		Simulation setting		Number		Simulated Power					
	Group of patients	Significance level	Hazard ratio	Required	Required	Expected Total	Expected	Expected	FSP	ESP	Power
				Number of events	Number of patients	study period (months)	Number of events	Number of patients			
25%	Biomarker-negative	0.0125	0.9	3425	3720	17.6	2014	2187	0.3678	0.1816	0.5678
	Biomarker-positive	0.0125	0.7	299	434	16.9	168	245	0.3924	0.1897	0.5392
	Entire population	0.025	-	3724	4154	-	2182	2432	-	-	-
	Biomarker-negative	0.015	0.9	3268	3550	16.8	1831	1989	0.4171	0.1691	0.4995
	Biomarker-positive	0.010	0.7	316	458	14	147	213	0.5076	0.2056	0.4245
	Entire population	0.025	-	3584	4008	-	1978	2202	-	-	-
	Biomarker-negative	0.010	0.9	3616	3928	14.9	1794	1948	0.4838	0.1882	0.4436

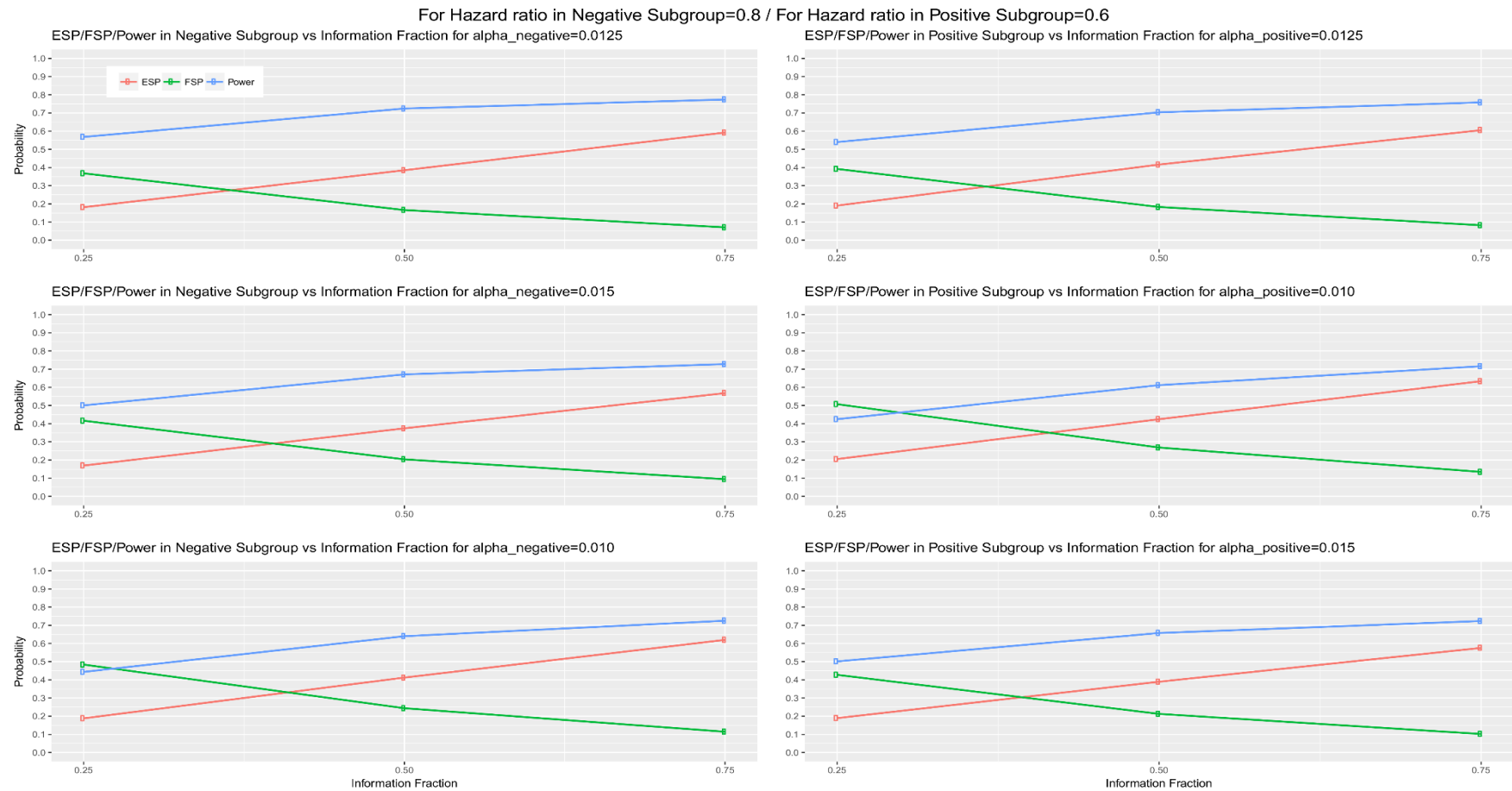


	Biomarker-positive	0.015	0.7	285	413	16.1	153	222	0.4277	0.1892	0.5010
	Entire population	0.025	-	3901	4341	-	1947	2170	-	-	-
50%	Biomarker-negative	0.0125	0.9	3425	3720	21.7	2482	2696	0.1656	0.3849	0.7248
	Biomarker-positive	0.0125	0.7	299	434	21	210	304	0.1835	0.4151	0.7017
	Entire population	0.025	-	3724	4154	-	2692	3000	-	-	-
	Biomarker-negative	0.015	0.9	3268	3550	21.3	2324	2524	0.2046	0.3732	0.6704
	Biomarker-positive	0.010	0.7	316	458	19.6	206	299	0.2681	0.4260	0.6129
	Entire population	0.025	-	3584	4008	-	2530	2823	-	-	-
	Biomarker-negative	0.010	0.9	3616	3928	20.2	2430	2639	0.2437	0.4124	0.6401
	Biomarker-positive	0.015	0.7	285	413	21.0	199	289	0.2127	0.3894	0.6575
	Entire population	0.025	-	3901	4341	-	2629	2928	-	-	-

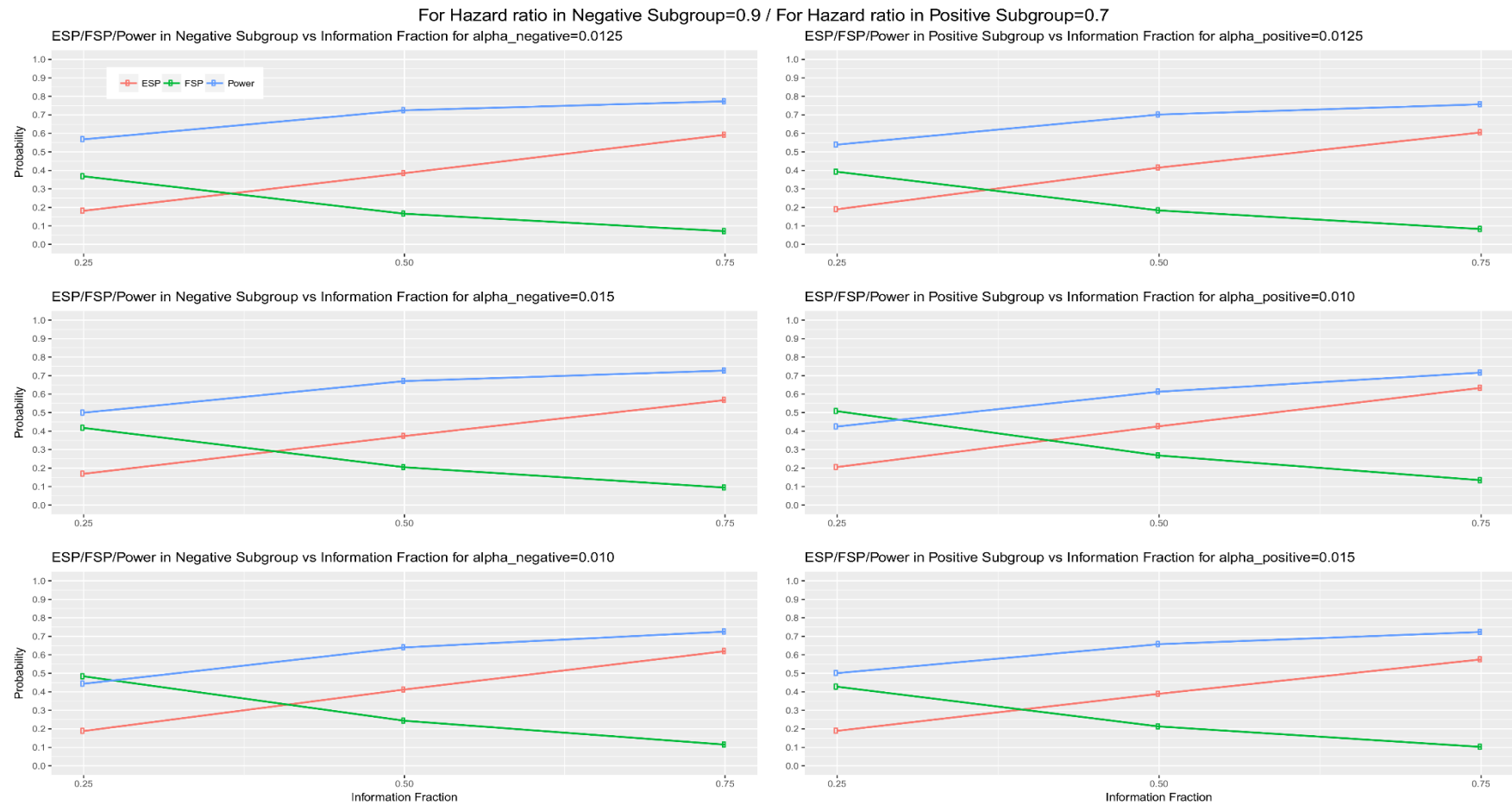
75%	Biomarker-negative	0.0125	0.9	3425	3720	25.0	2857	3103	0.0709	0.5925	0.7734
	Biomarker-positive	0.0125	0.7	299	434	24.8	248	359	0.0826	0.6051	0.7572
	Entire population	0.025	-	3724	4154	-	3105	3462	-	-	-
	Biomarker-negative	0.015	0.9	3268	3550	25.0	2727	2962	0.0944	0.5679	0.7272
	Biomarker-positive	0.010	0.7	316	458	24.2	255	370	0.1344	0.6338	0.7162
	Entire population	0.025	-	3584	4008	-	2982	3332	-	-	-
	Biomarker-negative	0.010	0.9	3616	3928	24.5	2952	3206	0.1146	0.6199	0.7254
	Biomarker-positive	0.015	0.7	285	413	24.9	237	343	0.1022	0.5752	0.7236
	Entire population	0.025	-	3901	4341	-	3189	3549	-	-	-



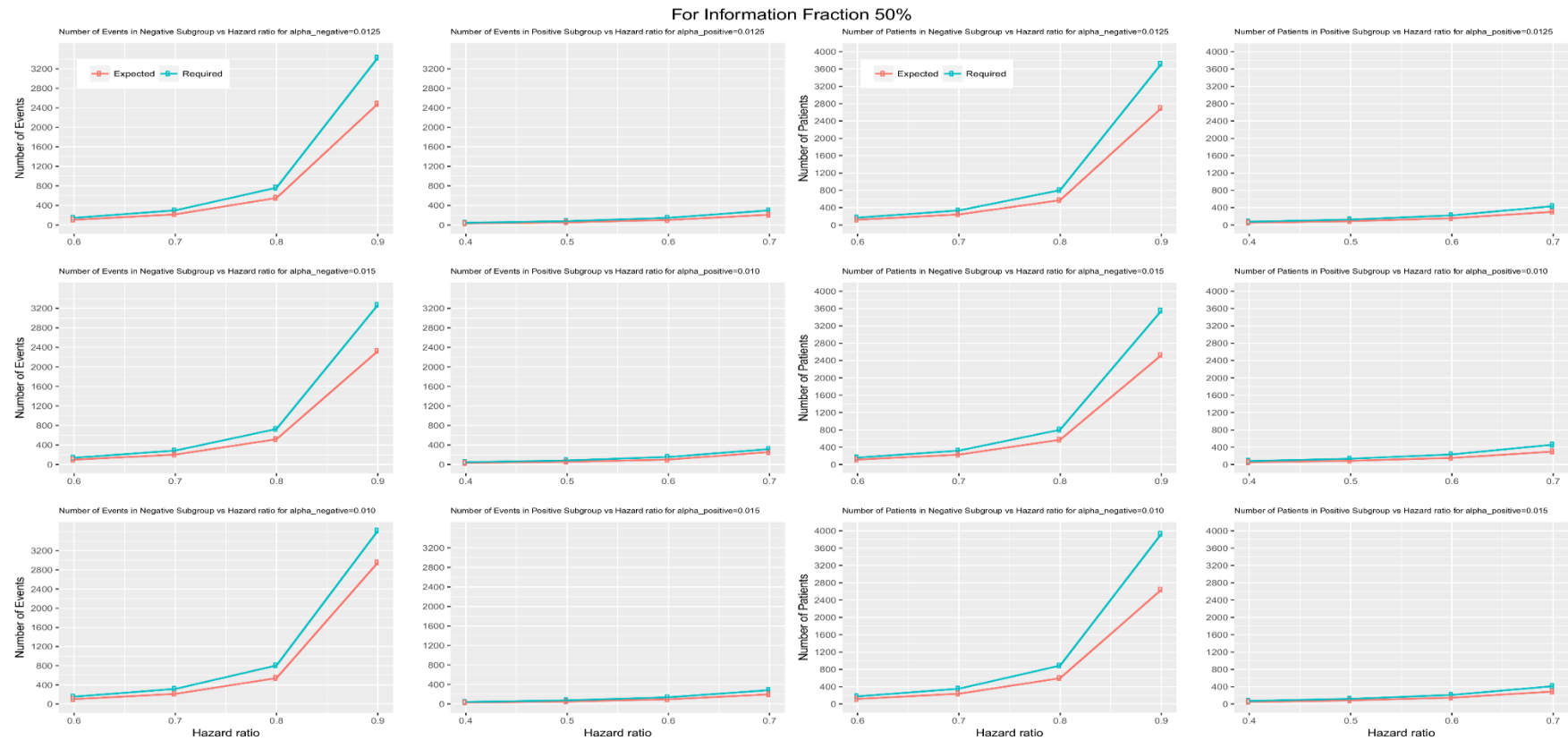
**Figure C.1.** Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 2 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively.



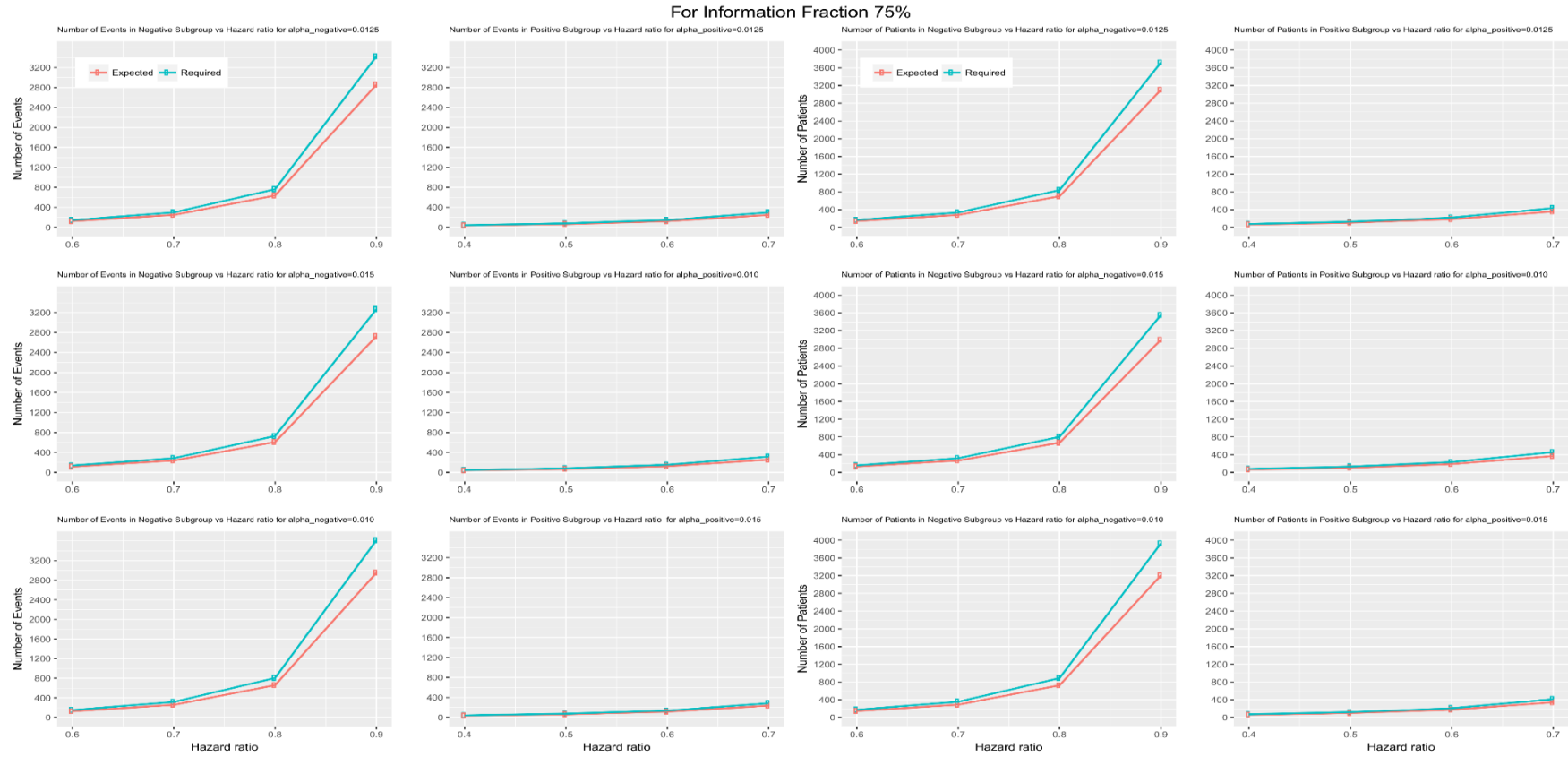
**Figure C.2.** Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 3 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i)  $\alpha_{-} = \alpha_{+} = 0.0125$ , (ii)  $\alpha_{-} = 0.015$  and  $\alpha_{+} = 0.010$  and (iii)  $\alpha_{-} = 0.010$  and  $\alpha_{+} = 0.015$  respectively.



**Figure C.3.** Efficacy stopping probability, futility stopping probability and power of a two-stage design versus the interim fraction (25%, 50%, 75%) in each biomarker-defined subgroup for scenario 4 of hazard ratios. Each row of graphs represents the different probabilities versus the interim fraction of each biomarker-defined subgroup when (i)  $\alpha_{-} = \alpha_{+} = 0.0125$ , (ii)  $\alpha_{-} = 0.015$  and  $\alpha_{+} = 0.010$  and (iii)  $\alpha_{-} = 0.010$  and  $\alpha_{+} = 0.015$  respectively.



**Figure C.4.** Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 50%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$ .



**Figure C.5.** Expected number of events and patients in two-stage design and required number of events and patients in one-stage design for each biomarker-defined subgroup versus the hazard ratios of each biomarker-defined subgroup when the interim fraction is 75%. The first two graphical representations in each row of graphs represent the number of events versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$  respectively. The remaining graphical representations in each row of graphs represent the number of patients versus the hazard ratio of each biomarker-defined subgroup when (i)  $\alpha_- = \alpha_+ = 0.0125$ , (ii)  $\alpha_- = 0.015$  and  $\alpha_+ = 0.010$  and (iii)  $\alpha_- = 0.010$  and  $\alpha_+ = 0.015$ .

## C.2. R codes

---

R codes for the Parallel Subgroup-Specific design and its adaptive version are given in C.2.1 and C.2.2 respectively. The R codes in C.2.2 are created based on examples and codes found in Chang M. Adaptive Design Theory and Implementation Using SAS and R, Second Edition. 2nd ed. London: CRC Press; 2014.

### C.2.1. Parallel Subgroup-Specific design

---

###The following codes correspond to case 1. Similar codes are ###used for the remaining cases (Please change the ###corresponding parameters)

#### **Case 1**

#### a1=0.0125, a2=0.0125  
#### hr.negative=0.7, hr.positive 0.5

med.control.negative =5  
med.control.positive=10

hr.negative = 0.7  
hr.positive=0.5

a<-0.025  
a2<-0.0125  
a1<-a-a2

number.of.events.positive<-(4\*((qnorm(a2)+qnorm(0.20))^2))/(log(hr.positive))^2  
number.of.events.positive  
number.of.events.negative<-(4\*((qnorm(a1)+qnorm(0.20))^2))/(log(hr.negative))^2  
number.of.events.negative

med.tmt.negative<- med.control.negative / hr.negative



```

med.tmt.negative
med.tmt.positive<- med.control.positive / hr.positive
med.tmt.positive

A.positive<-18
T.positive<-30
F.positive<-T.positive-A.positive

n.positive<-(number.of.events.positive)
n.positive
cut2<-round(n.positive)
cut2

pe.positive<-1-((1/((log(2)/ med.tmt.positive)*A.positive))*((exp(-(log(2) /
med.tmt.positive)*F.positive))-(exp(-(log(2) /
med.tmt.positive)*(A.positive+F.positive))))))
pc.positive<-1-((1/((log(2)/ med.control.positive)*A.positive))*((exp(-(log(2) /
med.control.positive)*F.positive))-(exp(-(log(2) /
med.control.positive)*(A.positive+F.positive))))))

N.positive<-n.positive/(0.5*pe.positive+0.5*pc.positive)
N.positive
n2<-round(N.positive)
n2
n13<-round(N.positive/2)
n13

A.negative<-18
T.negative<-30
F.negative<-T.negative-A.negative

n.negative<-(number.of.events.negative)
n.negative
cut1<-round(n.negative)
cut1

pe.negative<-1-((1/((log(2) / med.tmt.negative)*A.negative))*((exp(-(log(2) /
med.tmt.negative)*F.negative))-(exp(-(log(2) /
med.tmt.negative)*(A.negative+F.negative))))))

```

```
pc.negative<-1-((1/((log(2) / med.control.negative)*A.negative))*((exp(-(log(2) /
med.control.negative)*F.negative))-(exp(-(log(2) /
med.control.negative)*(A.negative+F.negative))))))
```

```
N.negative<-n.negative/(0.5*pe.negative+0.5*pc.negative)
N.negative
n1<-round(N.negative)
n1
n11<-round(N.negative/2)
n11
```

```
kappa<-0.5
p<-0.5
```

```
permonth.negative<-round(N.negative/A.negative)
permonth.negative
permonth.positive<-round(N.positive/A.positive)
permonth.positive
```

```
#### Fix enrolment and vary study duration
```

```
simCombTest <- function(n11=n11, n13=n13,
med.control.negative=med.control.negative,
med.control.positive=med.control.positive,
hr.negative=hr.negative,hr.positive=hr.positive,
permonth.negative=permonth.negative,
permonth.positive=permonth.positive, cut1=cut1,cut2=cut2, kappa=kappa, p=p){
```

```
# number of patients nij
# i = 1, 2 (stages) / j = control, tmt
```

```
### negative
n12 <- n11 * (1 - kappa) / kappa
n1 <- n11 + n12
```

```
#### positive
n14 <- n13 * (1 - kappa) / kappa
n2<-n13+n14
```

```
####total
```

```

n<-n1+n2

# generate arrival times for negative
arrivetime.negative <- NULL
for (i in 1:ceiling(n1 / permonth.negative)){arrivetime.negative <-
c(arrivetime.negative, runif(permonth.negative, min =i- 1, max =i ))}
arrivetime.negative <- arrivetime.negative[1:n1]

# generate arrival times for positive
arrivetime.positive <- NULL
for (i in 1:ceiling(n2 / permonth.positive)){arrivetime.positive <-
c(arrivetime.positive, runif(permonth.positive, min =i- 1, max =i ))}
arrivetime.positive <- arrivetime.positive[1:n2]

# generate event times for negative population
med.tmt.negative<- med.control.negative / hr.negative
time11 <- rexp(n11, rate = log(2) / med.control.negative)
time12 <- rexp(n12, rate = log(2) / med.tmt.negative)

# generate event times for positive population
med.tmt.positive<- med.control.positive / hr.positive
time13 <- rexp(n13, rate = log(2) / med.control.positive)
time14 <- rexp(n14, rate = log(2) / med.tmt.positive)

# observed times for stage 1 - negative patients
choose1 <- sample(1:(n11 + n12))
choose11 <- sort(choose1[1:n11])
choose12 <- sort(choose1[(n11 + 1):(n11 + n12)])

# observed times for stage 1 - positive patients
choose2 <- sample(1:(n13+ n14))
choose13 <- sort(choose2[1:n13])
choose14 <- sort(choose2[(n13 + 1):(n13 + n14)])

#####total time is event time plus arrival time

```

```

tottime11 <- arrivetime.negative[choose11] + time11 ##### #total study period for
those on control treatment who are #negative
tottime12 <- arrivetime.negative[choose12] + time12 ##### #total study period for
those on experimental treatment who #are negative
tottime1 <- c(tottime11, tottime12) ##### negative

tottime13 <- arrivetime.positive[choose13] + time13 ##### #total study period for
those on control treatment who are #positive
tottime14 <- arrivetime.positive[choose14] + time14 ##### #total study period for
those on experimental treatment who #are positive
tottime2 <- c(tottime13, tottime14) ##### positive

# find cutoff for stage 1-negative patients
cutoff1 <- sort(tottime1)[cut1] ### we mean the time after #study start at which a
prespecified number of events has #been reached
event1 <- ifelse(tottime1 > cutoff1, 0, 1)

# find cutoff for stage 1-positive patients
cutoff2 <- sort(tottime2)[cut2] ### we mean the time after study start at which a
prespecified number of events has been reached
event2 <- ifelse(tottime2 > cutoff2, 0, 1)

# apply censoring-negative
event.negative <- event1
time.negative <- c(time11, time12)

tmt.negative <- c(rep(0, n11), rep(1, n12))
stage.negative <- c(rep(1, n11 + n12))
subgroup.negative<-c(rep(0,n11 + n12))

# apply censoring-positive
event.positive <- event2
time.positive <- c(time13, time14)

tmt.positive <- c(rep(0, n13), rep(1, n14))

```

```

stage.positive <- c(rep(1, n13 + n14))
subgroup.positive<-c(rep(1,n13 + n14))

time<-c(time.negative,time.positive)
event<-c(event.negative,event.positive)
tmt<-c(tmt.negative,tmt.positive)
subgroup<-c(subgroup.negative,subgroup.positive)

dat.overall <- data.frame(time, event, tmt,subgroup)
colnames(dat.overall) <- c("time", "event", "tmt","subgroup")

data.frame(id=1:n,
           time=time,
           event=event,
           tmt=tmt,
           subgroup=subgroup,
           cutoff1=cutoff1,
           cutoff2=cutoff2)

}

set.seed(09032014)

nsim<-10000

hazratn1<-rep(NA,nsim)
hazratp1 <- rep(NA,nsim)
bettern1<- rep(NA,nsim)
betterp1<- rep(NA,nsim)
zn1<- rep(NA,nsim)
zp1<- rep(NA,nsim)

pn1<- rep(NA,nsim)
pp1<- rep(NA,nsim)
pnp1<- rep(NA,nsim)

cutoff1<-rep(NA,nsim)
cutoff2<-rep(NA,nsim)

```

```

for(k in 1:nsim){
  dat <- simCombTest(n11=n11, n13=n13,
med.control.negative=med.control.negative,
med.control.positive=med.control.positive,
hr.negative=hr.negative,hr.positive=hr.positive,
                    permonth.negative=permonth.negative,
permonth.positive=permonth.positive, cut1=cut1,cut2=cut2, kappa=kappa, p=p)

  cutoff1[k]<-dat$cutoff1[1]
  cutoff2[k]<-dat$cutoff2[1]

  # cox model in negative:
  ind1 <- (dat$subgroup== 0 & !is.na(dat$subgroup))
  if(sum(ind1) > 0){
    hazard1 <- coxph(Surv(time = time[ind1], event = event[ind1])
~tmt[ind1],data=dat)
    test1 <- survdiff(Surv(time =time[ind1], event = event[ind1]) ~ tmt[ind1],data=dat)
  } else {
    warning("Subgroup analysis not possible - all subgroup on same treatment")
    hazard2 <- NA
    hazratn1 <- NA
    pn1 <- NA
    zn1 <- NA
    test2 <- NA
  }

  # cox model in positive only:
  ind2 <- (dat$subgroup== 1 & !is.na(dat$subgroup))
  if(sum(ind2) > 0){
    hazard2 <- coxph(Surv(time = time[ind2], event = event[ind2])
~tmt[ind2],data=dat)
    test2 <- survdiff(Surv(time =time[ind2], event = event[ind2]) ~ tmt[ind2],data=dat)
  } else {
    warning("Subgroup analysis not possible - all subgroup on same treatment")
    hazard2 <- NA
    hazratp1 <- NA
    pp1 <- NA
    zp1 <- NA
    test2 <- NA
  }
}

```

```

# save hazard ratios
hazratn1[k] <- exp(hazard1$coef)
hazratp1[k] <- exp(hazard2$coef)

zn1[k]<- sqrt(test1$chi)
zp1[k]<- sqrt(test2$chi)

#### 1-sided

pn1[k]<-(2*(1 - pnorm(zn1[k])))/2
pp1[k]<-(2*(1 - pnorm(zp1[k])))/2

pnp1[k]<- min(2 * min(pn1[k], pp1[k]), max(pn1[k], pp1[k]))

#write.table(dat,file=" ",sep = ",",append=TRUE,dec=".",qmethod="double")

}

rejection.probability.negative<-sum( pn1<a1)/nsim
rejection.probability.positive<-sum( pp1<a2)/nsim

meanhazn1 <- exp(mean(log(hazratn1), na.rm = TRUE))
meanhazp1 <- exp(mean(log(hazratp1), na.rm = TRUE))
medhazn1 <- median(hazratn1, na.rm = TRUE)
medhazp1 <- median(hazratp1, na.rm = TRUE)
meanpn1 <- mean(pn1, na.rm = TRUE)
meanpp1 <- mean(pp1, na.rm = TRUE)
meancutoff1<-mean(cutoff1)
meancutoff2<-mean(cutoff2)

res <-
list("cut1"=cut1,"cut2"=cut2,"sumevents"=cut1+cut2,"n1"=n1,"n2"=n2,"sumsize"=n1+n
2,"a1"=a1,"a2"=a2,"HR negative" = meanhazn1,"HR positive" =meanhazp1,
      "rejection.probability.negative"=rejection.probability.negative,
      "rejection.probability.positive"=rejection.probability.positive,

```

```

"meancutoff1"=meancutoff1,"meancutoff2"=meancutoff2)

res

results<-unlist(cbind(res))
results<-data.frame((c("Number of events_Negative","Number of
events_Positive","total number of events","Sample size_Negative","Sample
size_Positive","total sample size",
"Alpha level_Negative","Alpha level_Positive","HR negative","HR
positive",
"rejection.probability.negative","rejection.probability.positive","Total
study period_Negative",
"Total study period_Positive")),round(results,3))

names(results)<-c("Definition","Results")

results

save(results, file = paste(path.results, "results_simulation_scenario=1", ".rdata", sep =
""))
save(results, file = paste(path.code, "code_simulation_scenario=1", ".rdata", sep = ""))

```

### C.2.2. An adaptive version of the Parallel Subgroup-Specific design

---

### The following codes correspond to Scenario 1. Similar codes are used for other scenarios (Please change the corresponding parameters).

```

#### Scenario 1
#### a1=0.0125, a2=0.0125
#### hr.negative=0.6, hr.positive 0.4
#### InfoTime.negative=0.25 (we will try also 0.50, 0.75)
#### InfoTime.positive=0.25 (we will try also 0.50, 0.75)

med.control.negative =5
med.control.positive=10

```



```

hr.negative = 0.6
hr.positive=0.4

a<-0.025
a2<-0.0125
a1<-a-a2

number.of.events.positive<-(4*((qnorm(a2)+qnorm(0.20))^2))/(log(hr.positive))^2
number.of.events.positive
number.of.events.negative<-(4*((qnorm(a1)+qnorm(0.20))^2))/(log(hr.negative))^2
number.of.events.negative

med.tmt.negative<- med.control.negative / hr.negative
med.tmt.negative
med.tmt.positive<- med.control.positive / hr.positive
med.tmt.positive

A.positive<-18
T.positive<-30
F.positive<-T.positive-A.positive

n.positive<-(number.of.events.positive)
n.positive<-round(n.positive)

pe.positive<-1-((1/((log(2)/ med.tmt.positive)*A.positive))*((exp(-(log(2) /
med.tmt.positive)*F.positive))-(exp(-(log(2) /
med.tmt.positive)*(A.positive+F.positive))))))
pc.positive<-1-((1/((log(2)/ med.control.positive)*A.positive))*((exp(-(log(2) /
med.control.positive)*F.positive))-(exp(-(log(2) /
med.control.positive)*(A.positive+F.positive))))))

N.positive<-n.positive/(0.5*pe.positive+0.5*pc.positive)
N.positive<-round(N.positive)

A.negative<-18
T.negative<-30
F.negative<-T.negative-A.negative

```

```
n.negative<-(number.of.events.negative)
n.negative<-round(n.negative)
```

```
pe.negative<-1-((1/((log(2) / med.tmt.negative)*A.negative))*((exp(-(log(2) /
med.tmt.negative)*F.negative))-(exp(-(log(2) /
med.tmt.negative)*(A.negative+F.negative))))))
pc.negative<-1-((1/((log(2) / med.control.negative)*A.negative))*((exp(-(log(2) /
med.control.negative)*F.negative))-(exp(-(log(2) /
med.control.negative)*(A.negative+F.negative))))))
```

```
N.negative<-n.negative/(0.5*pe.negative+0.5*pc.negative)
N.negative<-round(N.negative)
```

```
permonth.negative<-round(N.negative/A.negative)
permonth.negative
```

```
permonth.positive<-round(N.positive/A.positive)
permonth.positive
```

```
#### alpha1=early efficacy stopping boundary
#### beta1=early futility stopping boundary
### alpha2=final efficacy boundary
```

```
#log(hr.positive)
```

```
DCSPSurv2<-function(nSims=10000, Model="sum", alpha.positive=0.0125,
alpha.negative=0.0125, beta=0.2, tStd=30, tAcr=18,
lnHR.positive=log(hr.positive),lnHR.negative=log(hr.negative),
N.positive=N.positive, n.positive=n.positive, N.negative=N.negative,
n.negative=n.negative, InfoTime.positive=0.25,
InfoTime.negative=0.25,alpha1.positive=0.0080 , beta1.positive=0.1029,
alpha2.positive=0.1029,alpha1.negative=0.0070, beta1.negative=0.1129,
alpha2.negative=0.1129){
```

```
FSP.positive<-0;ESP.positive<-0;AveDs.positive<-0; Power.positive<-0;
FSP.negative<-0;ESP.negative<-0;AveDs.negative<-0; Power.negative<-0;
```

```

Ds.positive<-n.positive
Ds1.positive<-Ds.positive*InfoTime.positive

Ds.negative<-n.negative
Ds1.negative<-Ds.negative*InfoTime.negative


set.seed(09032014)


for(i in 1:nSims) {

  nFinal.positive<-Ds1.positive

  T1.positive<-rnorm(1)+sqrt(Ds1.positive/4)*(-lnHR.positive)

  nFinal.negative<-Ds1.negative

  T1.negative<-rnorm(1)+sqrt(Ds1.negative/4)*(-lnHR.negative)

  p1.positive<-1-pnorm(T1.positive)

  if (p1.positive>beta1.positive) {FSP.positive=FSP.positive+1/nSims}

  if (p1.positive<=alpha1.positive) {

    Power.positive=Power.positive+1/nSims; ESP.positive=ESP.positive+1/nSims

  }

  p1.negative<-1-pnorm(T1.negative)

  if (p1.negative>beta1.negative) {FSP.negative=FSP.negative+1/nSims}

  if (p1.negative<=alpha1.negative) {

```

```

Power.negative=Power.negative+1/nSims; ESP.negative=ESP.negative+1/nSims
}

if (p1.positive>alpha1.positive&p1.positive<=beta1.positive){

  nFinal.positive<-Ds.positive

  T2.positive<-rnorm(1)+sqrt((Ds.positive-Ds1.positive)/4)*(-lnHR.positive)

  p2.positive<-1-pnorm(T2.positive)

  if (Model=="sum"){TS2.positive=p1.positive+p2.positive}
  if (Model=="ind"){TS2.positive=p2.positive}
  if (Model=="prd"){TS2.positive=p1.positive*p2.positive}

  if (TS2.positive<=alpha2.positive) {Power.positive=Power.positive+1/nSims}
}

if (p1.negative>alpha1.negative&p1.negative<=beta1.negative){

  nFinal.negative<-Ds.negative

  T2.negative<-rnorm(1)+sqrt((Ds.negative-Ds1.negative)/4)*(-lnHR.negative)

  p2.negative<-1-pnorm(T2.negative)

  if (Model=="sum"){TS2.negative=p1.negative+p2.negative}
  if (Model=="ind"){TS2.negative=p2.negative}
  if (Model=="prd"){TS2.negative=p1.negative*p2.negative}

  if (TS2.negative<=alpha2.negative) {Power.negative=Power.negative+1/nSims}
}

```

```

AveDs.positive<-AveDs.positive+nFinal.positive/nSims
AveDs.negative<-AveDs.negative+nFinal.negative/nSims

}

#Padj=alpha1+Power-ESP

results=print(cbind(Model,
FSP.positive=round(FSP.positive,4),ESP.positive=round(ESP.positive,4),Power.positive=round(Power.positive,4),AveDs.positive=round(AveDs.positive),N.positive,

FSP.negative=round(FSP.negative,4),ESP.negative=round(ESP.negative,4),Power.negative=round(Power.negative,4),AveDs.negative=round(AveDs.negative),N.negative))

}

DCSPSurv2(nSims=10000, Model="sum", alpha.positive=0.0125,
alpha.negative=0.0125, beta=0.2, tStd=30, tAcr=18,
lnHR.positive=log(hr.positive),lnHR.negative=log(hr.negative),
N.positive=N.positive, n.positive=n.positive, N.negative=N.negative,
n.negative=n.negative, InfoTime.positive=0.25,
InfoTime.negative=0.25,alpha1.positive=0.0080 , beta1.positive=0.1029,
alpha2.positive=0.1029,alpha1.negative=0.0070, beta1.negative=0.1129,
alpha2.negative=0.1129)

###expected sample size
# calculate the following: (FSP+ESP)*N.positive
#N.positive*(ESP.positive+FSP.positive)

###expected duration
#(FSP.+ESP)*time from the first patient-in to the kth interim analysis
#(InfoTime*30)*(0.1731+0.7023)

```

## Appendix D

---

Appendix D includes supporting information related to Chapter 7. Results of Simulation study 1 are presented in Appendix D.1 and results of Simulation study 2 are given in Appendix D.2. R codes are provided in Appendix D.3.

## D.1. Results for Simulation study 1

**Table D.1.1.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method and O'Brien-Fleming decision boundaries.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
156	0.15	25%	Null	231	0.001930	0.024706
			Alternative	191	0.067743	0.870993
		50%	Null	246	0.001930	0.024706
			Alternative	182	0.179273	0.903351
		75%	Null	251	0.001930	0.024706
			Alternative	182	0.315503	0.907162

**Table D.1.2.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method and O'Brien-Fleming decision boundaries.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
156	0.15	25%	Null	233	0.001930	0.024706
			Alternative	150	0.067743	0.785080
		50%	Null	243	0.001930	0.024706
			Alternative	148	0.179273	0.819279
		75%	Null	250	0.001930	0.024706
			Alternative	162	0.315503	0.849604



**Table D.1.3.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method and Pocock decision boundaries.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
156	0.15	25%	Null	230	0.016078	0.021698
			Alternative	172	0.224221	0.761714
		50%	Null	246	0.016078	0.021698
			Alternative	163	0.426191	0.836597
		75%	Null	251	0.016078	0.021698
			Alternative	169	0.600358	0.886253

**Table D.1.4.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method and Pocock decision boundaries.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
156	0.15	25%	Null	246	0.016078	0.021698
			Alternative	176	0.224221	0.771788
		50%	Null	250	0.016078	0.021698
			Alternative	163	0.426191	0.838188
		75%	Null	253	0.016078	0.021698
			Alternative	167	0.600358	0.880163

**Table D.1.5.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method and O'Brien-Fleming decision boundaries.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
183	186	0.746	25%	Null	271	0.001930	0.024706
				Alternative	220	0.068805	0.863041
			50%	Null	282	0.001930	0.024706
				Alternative	211	0.182171	0.896072
			75%	Null	284	0.001930	0.024706
				Alternative	211	0.320094	0.899924
			25%	Null	643	0.001930	0.024706
				Alternative	576	0.069643	0.807397

550	558	0.845	50%	Null	654	0.001930	0.024706
				Alternative	553	0.184371	0.841650
			75%	Null	657	0.001930	0.024706
				Alternative	559	0.323890	0.837346
341	346	0.807	25%	Null	431	0.001930	0.024706
				Alternative	374	0.069385	0.827455
			50%	Null	442	0.001930	0.024706
				Alternative	359	0.183666	0.862614
			75%	Null	445	0.001930	0.024706
				Alternative	362	0.322716	0.862878
			25%	Null	308	0.001930	0.024706

220	224	0.765		Alternative	256	0.069015	0.851771
			50%	Null	319	0.001930	0.024706
				Alternative	246	0.182768	0.885632
			75%	Null	322	0.001930	0.024706
				Alternative	247	0.321049	0.889006

**Table D.1.6.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method and O'Brien-Fleming decision boundaries.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
183	186	0.746	25%	Null	241	0.001930	0.024706
				Alternative	174	0.068805	0.779716

				50%	Null	261	0.001930	0.024706
					Alternative	173	0.182171	0.815679
					75%	Null	275	0.001930
					Alternative	190	0.320094	0.846261
550	558	0.845	25%		Null	616	0.001930	0.024706
				Alternative	481	0.069643	0.740230	
				50%	Null	642	0.001930	0.024706
				Alternative	481	0.184371	0.782292	
				75%	Null	654	0.001930	0.024706
				Alternative	534	0.323890	0.807523	
	25%	Null		406	0.001930	0.024706		

341	346	0.807		Alternative	303	0.069385	0.755569
			50%	Null	429	0.001930	0.024706
				Alternative	307	0.183666	0.796379
			75%	Null	440	0.001930	0.024706
				Alternative	340	0.322716	0.825099
220	224	0.765	25%	Null	281	0.001930	0.024706
				Alternative	205	0.069015	0.772559
			50%	Null	302	0.001930	0.024706
				Alternative	205	0.182768	0.810128
			75%	Null	315	0.001930	0.024706
				Alternative	226	0.321049	0.840671

**Table D.1.7.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with effect-size ratio method and Pocock decision boundaries.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
183	186	0.746	25%	Null	269	0.016078	0.021698
				Alternative	198	0.226595	0.751778
			50%	Null	280	0.016078	0.021698
				Alternative	188	0.430316	0.826512
			75%	Null	283	0.016078	0.021698
				Alternative	195	0.605391	0.875019
550	558	0.845	25%	Null	637	0.016078	0.021698
				Alternative	510	0.228383	0.687093



			50%	Null	650	0.016078	0.021698
				Alternative	484	0.433748	0.75865
			75%	Null	654	0.016078	0.021698
				Alternative	512	0.609326	0.797284
341                      346                      0.807			25%	Null	427	0.016078	0.021698
				Alternative	333	0.227844	0.709340
			50%	Null	439	0.016078	0.021698
				Alternative	316	0.432635	0.783473
			75%	Null	442	0.016078	0.021698
				Alternative	333	0.608126	0.826463
			25%	Null	306	0.016078	0.021698

220	224	0.765		Alternative	230	0.226992	0.737624
			50%	Null	318	0.016078	0.021698
				Alternative	218	0.431123	0.812622
			75%	Null	320	0.016078	0.021698
				Alternative	228	0.606380	0.859731

**Table D.1.8.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method and Pocock decision boundaries.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Rejection probability
			25%	Null	263	0.016078	0.021698
				Alternative	202	0.226595	0.760786

183	186	0.746	50%	Null	275	0.016078	0.021698
				Alternative	188	0.430316	0.829126
			75%	Null	281	0.016078	0.021698
				Alternative	193	0.605391	0.871070
550	558	0.845	25%	Null	640	0.016078	0.021698
				Alternative	514	0.228383	0.691205
			50%	Null	650	0.016078	0.021698
				Alternative	489	0.433748	0.764274
			75%	Null	654	0.016078	0.021698
				Alternative	512	0.609326	0.797284
			25%	Null	428	0.016078	0.021698

341	346	0.807		Alternative	338	0.227844	0.716711
			50%	Null	438	0.016078	0.021698
				Alternative	320	0.432635	0.789683
			75%	Null	443	0.016078	0.021698
				Alternative	333	0.608126	0.826491
220	224	0.765	25%	Null	303	0.016078	0.021698
				Alternative	234	0.226992	0.746794
			50%	Null	314	0.016078	0.021698
				Alternative	220	0.431123	0.817145
			75%	Null	320	0.016078	0.021698
				Alternative	227	0.606380	0.858074

## D.2. Results for Simulation study 2

**Table D.2.1.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
156	0.15	25%	Null	54	0.00195	0.875893	0.018734
			Alternative	103	0.067891	0.408571	0.550792
		50%	Null	95	0.00195	0.875893	0.018734
			Alternative	145	0.179291	0.210752	0.744654
		75%	Null	132	0.00195	0.875893	0.018734
			Alternative	167	0.315418	0.107176	0.834594

**Table D.2.2.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
156	0.15	25%	Null	50	0.00195	0.875893	0.018734
			Alternative	74	0.067891	0.408571	0.486881
		50%	Null	89	0.00195	0.875893	0.018734
			Alternative	114	0.179291	0.210752	0.667259
		75%	Null	128	0.00195	0.875893	0.018734
			Alternative	147	0.315418	0.107176	0.777358

**Table D.2.3.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the effect-size ratio method, Pocock efficacy boundaries and Pocock futility boundary.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
156	0.15	25%	Null	53	0.015956	0.875933	0.020321
			Alternative	85	0.224141	0.408891	0.506438
		50%	Null	94	0.015956	0.875933	0.020321
			Alternative	126	0.425881	0.211056	0.714582
		75%	Null	131	0.015956	0.875933	0.020321
			Alternative	154	0.600368	0.107371	0.834428

**Table D.2.4.** Adaptive sample size re-estimation design for binary outcome applied to the reverse marker-based strategy design with the conditional power method, Pocock efficacy boundaries and Pocock futility boundary.

Fixed sample size	Difference in response rates between strategy arms	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
156	0.15	25%	Null	56	0.015956	0.875933	0.020321
			Alternative	88	0.224141	0.408891	0.517662
		50%	Null	95	0.015956	0.875933	0.020321
			Alternative	125	0.425881	0.211056	0.716136
		75%	Null	131	0.015956	0.875933	0.020321
			Alternative	152	0.600368	0.107371	0.828347



**Table D.2.5.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
183	186	0.746	25%	Null	68	0.001950	0.875893	0.018734
				Alternative	123	0.068890	0.405581	0.554985
			50%	Null	115	0.001950	0.875893	0.018734
				Alternative	171	0.182105	0.207649	0.746814
			75%	Null	157	0.001950	0.875893	0.018734
				Alternative	196	0.319987	0.104713	0.833836
550	558	0.845	25%	Null	196	0.001950	0.875893	0.018734
				Alternative	367	0.069634	0.403354	0.556305

			50%	Null	324	0.001950	0.875893	0.018734
				Alternative	476	0.184386	0.205221	0.730834
			75%	Null	447	0.001950	0.875893	0.018734
				Alternative	534	0.323629	0.102877	0.796348
341	346	0.807	25%	Null	123	0.001950	0.875893	0.018734
				Alternative	229	0.069406	0.404002	0.556719
			50%	Null	205	0.001950	0.875893	0.018734
				Alternative	303	0.183680	0.205953	0.738352
			75%	Null	282	0.001950	0.875893	0.018734
				Alternative	343	0.322485	0.103446	0.813330
			25%	Null	81	0.001950	0.875893	0.018734

220	224	0.765		Alternative	148	0.069069	0.405015	0.555808
			50%	Null	136	0.001950	0.875893	0.018734
				Alternative	202	0.182708	0.207027	0.744688
			75%	Null	186	0.001950	0.875893	0.018734
				Alternative	231	0.320885	0.104289	0.828374

**Table D.2.6.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method, O'Brien-Fleming efficacy boundaries and Pocock futility boundary.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
183	186	0.746	25%	Null	56	0.001950	0.875893	0.018734
				Alternative	88	0.068890	0.405581	0.489414

			50%	Null	102	0.001950	0.875893	0.018734
				Alternative	135	0.182105	0.207649	0.670295
			75%	Null	148	0.001950	0.875893	0.018734
				Alternative	175	0.319987	0.104713	0.780193
550                      558                      0.845			25%	Null	170	0.001950	0.875893	0.018734
				Alternative	264	0.069634	0.403354	0.491318
			50%	Null	310	0.001950	0.875893	0.018734
				Alternative	403	0.184386	0.205221	0.671262
			75%	Null	444	0.001950	0.875893	0.018734
				Alternative	509	0.323629	0.102877	0.766529
			25%	Null	105	0.001950	0.875893	0.018734

341	346	0.807		Alternative	164	0.069406	0.404002	0.490766
			50%	Null	191	0.001950	0.875893	0.018734
				Alternative	251	0.183680	0.205953	0.671945
			75%	Null	277	0.001950	0.875893	0.018734
				Alternative	321	0.322485	0.103446	0.775550
			220	224	0.765	25%	Null	67
Alternative	106	0.069069					0.405015	0.489909
50%	Null	123				0.001950	0.875893	0.018734
	Alternative	162				0.182708	0.207027	0.670911
75%	Null	179				0.001950	0.875893	0.018734
	Alternative	210				0.320885	0.104289	0.780054

**Table D.2.7.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the effect-size ratio method, Pocock efficacy boundaries and Pocock futility boundary.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
183	186	0.746	25%	Null	67	0.015956	0.875933	0.020321
				Alternative	101	0.226404	0.405947	0.511271
			50%	Null	113	0.015956	0.875933	0.020321
				Alternative	147	0.429997	0.207989	0.715421
			75%	Null	155	0.015956	0.875933	0.020321
				Alternative	180	0.605401	0.104909	0.829909
550	558	0.845	25%	Null	190	0.015956	0.875933	0.020321
				Alternative	301	0.228085	0.403724	0.511912

			50%	Null	320	0.015956	0.875933	0.020321
				Alternative	406	0.433466	0.205600	0.687413
			75%	Null	444	0.015956	0.875933	0.020321
				Alternative	487	0.609359	0.103043	0.774560
341	346	0.807	25%	Null	120	0.015956	0.875933	0.020321
				Alternative	188	0.227524	0.404389	0.512879
			50%	Null	202	0.015956	0.875933	0.020321
				Alternative	260	0.432332	0.206358	0.699500
			75%	Null	280	0.015956	0.875933	0.020321
				Alternative	313	0.608132	0.103625	0.796938
			25%	Null	79	0.015956	0.875933	0.020321

220	224	0.765		Alternative	122	0.226785	0.405382	0.512214
			50%	Null	134	0.015956	0.875933	0.020321
				Alternative	174	0.430812	0.207408	0.711023
			75%	Null	184	0.015956	0.875933	0.020321
				Alternative	212	0.606405	0.104468	0.820300

**Table D.2.8.** Adaptive sample size re-estimation design for time-to-event outcome applied to the reverse marker-based strategy design with the conditional power method, Pocock efficacy boundaries and Pocock futility boundary.

Fixed number of events	Fixed sample size	Hazard ratio	Information fraction	Hypothesis	Adaptive Average sample size	Efficacy stopping probability	Futility stopping probability	Rejection probability
			25%	Null	60	0.015956	0.875933	0.020321
				Alternative	105	0.226404	0.405947	0.520468



183	186	0.746	50%	Null	107	0.015956	0.875933	0.020321
				Alternative	148	0.429997	0.207989	0.717937
			75%	Null	153	0.015956	0.875933	0.020321
				Alternative	178	0.605401	0.104909	0.825912
550	558	0.845	25%	Null	186	0.015956	0.875933	0.020321
				Alternative	305	0.228085	0.403724	0.515823
			50%	Null	318	0.015956	0.875933	0.020321
				Alternative	412	0.433466	0.205600	0.693019
			75%	Null	445	0.015956	0.875933	0.020321
				Alternative	487	0.609359	0.103043	0.77456
			25%	Null	114	0.015956	0.875933	0.020321

341	346	0.807		Alternative	193	0.227524	0.404389	0.519999
			50%	Null	199	0.015956	0.875933	0.020321
				Alternative	264	0.432332	0.206358	0.705671
			75%	Null	279	0.015956	0.875933	0.020321
				Alternative	313	0.608132	0.103625	0.796938
220	224	0.765	25%	Null	73	0.015956	0.875933	0.020321
				Alternative	126	0.226785	0.405382	0.521006
			50%	Null	129	0.015956	0.875933	0.020321
				Alternative	176	0.430812	0.207408	0.715495
			75%	Null	183	0.015956	0.875933	0.020321
				Alternative	211	0.606405	0.104468	0.818615

## D.3. R codes

---

R codes for Simulation Study 1 and Simulation Study 2 are presented in D.3.1. and D.3.2 respectively. The following R codes are created based on examples and codes found in:

1. Chang M. Adaptive Design Theory and Implementation Using SAS and R, Second Edition. 2nd ed. London: CRC Press; 2014.
2. Francesca G, Luigi M. Two-stage re-estimation adaptive design: a simulation study. 2013; doi:10.2427/8862.

In D.3.3 we provide the R codes for the type I error probabilities derived with the O'Brien-Fleming and Pocock method.

### D.3.1. Simulation study 1

---

**#Binary Endpoint**

**#O'Brien-Fleming efficacy boundaries**

**# Conditional power method**

```
cp=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,  
            ry=0.57,n1=78,diff=0.15,alpha1=0.0026,alpha0=1,alpha2=0.0240,  
            w=1/sqrt(2),cP=0.8){
```

```
set.seed(1736)
```

```
FSP=0; ESP=0; nmean=0; power=0
```

```
nclassic=((((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-  
ry_0)))/((diff)^2)
```

```
nmax=nclassic+100
```

```
r=(rx+ry)/2
```

```
sigma=sqrt(r*(1-r))
```

```
for(iSim in 1:nsim){
```

```
  rx1=rnorm(1,rx,sigma/sqrt(n1))
```

```
  ry1=rnorm(1,ry,sigma/sqrt(n1))
```

```
  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
```

```
  p1=1-pnorm(t1)
```

```
  if (p1>alpha0){
```

```
    FSP=FSP+1/nsim
```

```
    n2=0
```

```
  }
```

```
  if (p1<=alpha1){
```

```
    power=power+1/nsim
```

```
    ESP=ESP+1/nsim
```

```
    n2=0
```

```
  }
```

```
  if (p1>alpha1 & p1<=alpha0) {
```

```
    eSize=diff/sigma
```

```
    Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w
```

```

n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2)

rx2=rnorm(1,rx,sigma/sqrt(n2))
ry2=rnorm(1,ry,sigma/sqrt(n2))

t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

z2=(t1+t2)/sqrt(2)

p2=1-pnorm(z2)

if (p2<=alpha2) {power=power+1/nsim}

}

nmean=nmean+(n1+n2)/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For rx=0.74 and ry=0.57

```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.74,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

### **#Binary Endpoint**

### **#O'Brien-Fleming efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=78, n0=156, diff=0.15,alpha1=0.0026, alpha0=1,
  alpha2=0.0240){
```

```
  set.seed(1736)
```

```
  FSP=0; ESP=0; nmean=0; power=0
```

```
  nclassic=((((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-
  ry_0)))/((diff)^2)
```

```
  nmax=nclassic+100
```

```
  r=(rx+ry)/2
```

```
  sigma=sqrt(r*(1-r))
```

```

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(n0, (er^2)*n0))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

      ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

      t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

      z2=(t1+t2)/sqrt(2)

      p2=1-pnorm(z2)

      if (p2<=alpha2) {power=power+1/nsim}

    }

  }

  nmean=nmean+nfinal/nsim

```

```

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.57,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.002, alpha0=1,
    alpha2=0.0240)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.002, alpha0=1,
    alpha2=0.0240)

#### Similar codes for 25% and 75% Information Fraction

#Survival Endpoint

#O'Brien-Fleming efficacy boundaries

# Conditional power method

cp=function(nsim=1000000,case="null",alpha=0.025, beta=0.2,
    ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
    f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
    w=0.70711,cP=0.9){

set.seed(1736)

```



```
FSP=0; ESP=0; Nmean=0; power=0
```

```
m.strategy=(k*ma.pos+(1-k)*mb.neg)  
m.reverse=(k*mb.pos+(1-k)*ma.neg)
```

```
nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)
```

```
hr=m.strategy/m.reverse
```

```
prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-  
tAcr))-exp((-log(2)/m.strategy)*tStd)))
```

```
prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-tAcr))-  
exp((-log(2)/m.reverse)*tStd)))
```

```
Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)
```

```
n1=Nclassic*f
```

```
nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients
```

```
rx=log(2)/m.strategy  
ry=log(2)/m.reverse
```

```
diff=rx-ry
```

```
if (case=="null"){rx=ry}
```

```
r=(rx+ry)/2
```

```
expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)
```

```
sigma=r*(1+expTerm)^(-0.5)
```

```
for(iSim in 1:nsim){
```

```

rx1=rnorm(1,rx,sigma/sqrt(n1))

ry1=rnorm(1,ry,sigma/sqrt(n1))

t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

p1=1-pnorm(t1)

if (p1>alpha0){

  FSP=FSP+1/nsim
  n2=0
}

if (p1<=alpha1){

  power=power+1/nsim
  ESP=ESP+1/nsim
  n2=0
}

if (p1>alpha1 & p1<=alpha0) {

  eSize=diff/sigma

  Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

  n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

  rx2=rnorm(1,rx,sigma/sqrt(n2))
  ry2=rnorm(1,ry,sigma/sqrt(n2))

  t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

  z2=(t1+t2)/sqrt(2);

  p2=1-pnorm(z2);

  if (p2<=alpha2) {power=power+1/nsim}

}

Nmean=Nmean+(n1+n2)/nsim

```

```

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.024,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75%

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75%

```

```
## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75%
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% information fraction
```

### **#Survival Endpoint**

### **#O'Brien-Fleming efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,  
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
```

```

w=0.70711,cP=0.9){

set.seed(1736)

FSP=0; ESP=0; Nmean=0; power=0

m.strategy=(k*ma.pos+(1-k)*mb.neg)
m.reverse=(k*mb.pos+(1-k)*ma.neg)

nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)

hr=m.strategy/m.reverse

prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-
tAcr))-exp((-log(2)/m.strategy)*tStd)))

prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-
tAcr))-exp((-log(2)/m.reverse)*tStd)))

Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)

n1=Nclassic*f

nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients

rx=log(2)/m.strategy
ry=log(2)/m.reverse

diff=rx-ry

if (case=="null"){rx=ry}

r=(rx+ry)/2

expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)

sigma=r*(1+expTerm)^(-0.5)

```

```

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(Nclassic, (er^2)*Nclassic))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

      ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

      t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

      z2=(t1+t2)/sqrt(2)

      p2=1-pnorm(z2)

      if (p2<=alpha2) {power=power+1/nsim}

    }

  }

  Nmean=Nmean+nfinal/nsim

}

```

```

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% information fraction

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% information fraction

```

## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,  
mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,  
mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

#### Similar codes for 50% and 75% information fraction

## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,  
mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,  
mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=1,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

#### Similar codes for 50% and 75% information fraction



**#Binary Endpoint**

**#Pocock efficacy boundaries**

**# Conditional power method**

```
cp=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,  
            ry=0.57,n1=78,diff=0.15,alpha1=0.0026,alpha0=1,alpha2=0.0240,  
            w=1/sqrt(2),cP=0.8){
```

```
  set.seed(1736)
```

```
  FSP=0; ESP=0; nmean=0; power=0
```

```
  nclassic=((qnorm(1-alpha)+ qnorm(1-beta))^2*(rx_0*(1-rx_0)+ry_0*(1-  
    ry_0)))/((diff)^2)
```

```
  nmax=nclassic+100
```

```
  r=(rx+ry)/2
```

```
  sigma=sqrt(r*(1-r))
```

```
  for(iSim in 1:nsim){
```

```
    rx1=rnorm(1,rx,sigma/sqrt(n1))
```

```
    ry1=rnorm(1,ry,sigma/sqrt(n1))
```

```
    t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
```

```
    p1=1-pnorm(t1)
```

```

if (p1>alpha0){

  FSP=FSP+1/nsim
  n2=0
}

if (p1<=alpha1){

  power=power+1/nsim
  ESP=ESP+1/nsim
  n2=0
}

if (p1>alpha1 & p1<=alpha0) {

  eSize=diff/sigma

  Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

  n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

  rx2=rnorm(1,rx,sigma/sqrt(n2))
  ry2=rnorm(1,ry,sigma/sqrt(n2))

  t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

  z2=(t1+t2)/sqrt(2);

  p2=1-pnorm(z2);

  if (p2<=alpha2) {power=power+1/nsim}

}

nmean=nmean+(n1+n2)/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=1,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=1,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)
```

```
#### Similar codes for 50% and 75% information fraction
```

```
## For rx=0.74 and ry=0.57
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=1,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.74,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=1,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)
```

```
#### Similar codes for 50% and 75% information fraction
```

### **#Binary Endpoint**

### **#Pocock efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=78, n0=156, diff=0.15,alpha1=0.0026, alpha0=1,
  alpha2=0.0240){
```

```

set.seed(1736)

FSP=0; ESP=0; nmean=0; power=0

nclassic=((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-
ry_0)))/((diff)^2)

nmax=nclassic+100

r=(rx+ry)/2

sigma=sqrt(r*(1-r))

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(n0, (er^2)*n0))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

```

```

    ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

    t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

    z2=(t1+t2)/sqrt(2)

    p2=1-pnorm(z2)

    if (p2<=alpha2) {power=power+1/nsim}

  }

}

nmean=nmean+nfinal/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.57,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.016, alpha0=1,
    alpha2=0.009)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.016, alpha0=1,
    alpha2=0.009)

#### Similar codes for 50% and 75% Information fraction

```

**#Survival Endpoint**

**#Pocock efficacy boundaries**

**# Conditional power method**

```
cp=function(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
w=0.70711,cP=0.9){

set.seed(1736)

FSP=0; ESP=0; Nmean=0; power=0

m.strategy=(k*ma.pos+(1-k)*mb.neg)
m.reverse=(k*mb.pos+(1-k)*ma.neg)

nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)

hr=m.strategy/m.reverse

prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-
tAcr))-exp((-log(2)/m.strategy)*tStd)))

prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-tAcr))-
exp((-log(2)/m.reverse)*tStd)))

Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)

n1=Nclassic*f

nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients
```

```

rx=log(2)/m.strategy
ry=log(2)/m.reverse

diff=rx-ry

if (case=="null"){rx=ry}

r=(rx+ry)/2

expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)

sigma=r*(1+expTerm)^(-0.5)


for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){

    FSP=FSP+1/nsim
    n2=0
  }

  if (p1<=alpha1){

    power=power+1/nsim
    ESP=ESP+1/nsim
    n2=0
  }

  if (p1>alpha1 & p1<=alpha0) {

    eSize=diff/sigma

```

```

Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

rx2=rnorm(1,rx,sigma/sqrt(n2))
ry2=rnorm(1,ry,sigma/sqrt(n2))

t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

z2=(t1+t2)/sqrt(2);

p2=1-pnorm(z2);

if (p2<=alpha2) {power=power+1/nsim}

}

Nmean=Nmean+(n1+n2)/nsim

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

```



```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
#Survival Endpoint
```

```
#Pocock efficacy boundaries
```

```
# Effect size ratio method
```

```
es=function(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,  
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,  
w=0.70711,cP=0.9){  
  
  set.seed(1736)  
  
  FSP=0; ESP=0; Nmean=0; power=0  
  
  m.strategy=(k*ma.pos+(1-k)*mb.neg)  
  m.reverse=(k*mb.pos+(1-k)*ma.neg)  
  
  nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)  
  
  hr=m.strategy/m.reverse  
  
  prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-  
tAcr))-exp((-log(2)/m.strategy)*tStd)))  
  
  prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-  
tAcr))-exp((-log(2)/m.reverse)*tStd)))  
  
  Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)  
  
  n1=Nclassic*f  
  
  nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients
```

```

rx=log(2)/m.strategy
ry=log(2)/m.reverse

diff=rx-ry

if (case=="null"){rx=ry}

r=(rx+ry)/2

expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)

sigma=r*(1+expTerm)^(-0.5)


for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(Nclassic, (er^2)*Nclassic))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

      ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

```

```

t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

z2=(t1+t2)/sqrt(2)

p2=1-pnorm(z2)

if (p2<=alpha2) {power=power+1/nsim}

}

}

Nmean=Nmean+nfinal/nsim

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping

```

```

es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping

```

```
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=1,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

### Similar codes for 505 and 75% Information fraction

### D.3.2. Simulation study 2

---

**#Binary Endpoint**

**#O' Brien-Fleming efficacy boundaries**

**# Conditional power method**

```
cp=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=78,diff=0.15,alpha1=0.0026,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8){
```

```
  set.seed(1736)
```

```
  FSP=0; ESP=0; nmean=0; power=0
```

```
  nclassic=((((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-
  ry_0)))/((diff)^2)
```

```
  nmax=nclassic+100
```

```
  r=(rx+ry)/2
```

```
  sigma=sqrt(r*(1-r))
```

```

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))
  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
  p1=1-pnorm(t1)

  if (p1>alpha0){

    FSP=FSP+1/nsim
    n2=0

  }

  if (p1<=alpha1){

    power=power+1/nsim
    ESP=ESP+1/nsim
    n2=0

  }

  if (p1>alpha1 & p1<=alpha0) {

    eSize=diff/sigma

    Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

    n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2)

    rx2=rnorm(1,rx,sigma/sqrt(n2))
    ry2=rnorm(1,ry,sigma/sqrt(n2))

    t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

    z2=(t1+t2)/sqrt(2)

```

```

    p2=1-pnorm(z2)

    if (p2<=alpha2) {power=power+1/nsim}

  }

  nmean=nmean+(n1+n2)/nsim
}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=0.124,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=0.124,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For rx=0.74 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=0.124,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.74,
  ry=0.57,n1=39,diff=0.15,alpha1=0.002,alpha0=0.124,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8)

```



```
#### Similar codes for 50% and 75% Information fraction
```

```
#Binary Endpoint
```

```
#O' Brien-Fleming efficacy boundaries
```

```
# Effect size ratio method
```

```
es=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,  
            ry=0.57,n1=78, n0=156, diff=0.15,alpha1=0.0026, alpha0=1,  
            alpha2=0.0240){
```

```
  set.seed(1736)
```

```
  FSP=0; ESP=0; nmean=0; power=0
```

```
  nclassic=((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-  
ry_0)))/((diff)^2)
```

```
  nmax=nclassic+100
```

```
  r=(rx+ry)/2
```

```
  sigma=sqrt(r*(1-r))
```

```
  for(iSim in 1:nsim){
```

```
    rx1=rnorm(1,rx,sigma/sqrt(n1))
```

```
    ry1=rnorm(1,ry,sigma/sqrt(n1))
```

```
    t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
```

```

p1=1-pnorm(t1)

if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

if (p1>alpha1 & p1<=alpha0) {

  if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

  er=diff/(abs(rx1-ry1)+0.0000001)

  nfinal=min(nmax, max(n0, (er^2)*n0))

  if (nfinal>n1) {

    rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

    ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

    t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

    z2=(t1+t2)/sqrt(2)

    p2=1-pnorm(z2)

    if (p2<=alpha2) {power=power+1/nsim}

  }

}

nmean=nmean+nfinal/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.57,

```

```
ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.002, alpha0=0.124,
alpha2=0.0240)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.002, alpha0=0.124,
alpha2=0.0240)
```

```
#### Similar codes for 50% and 75% Information fraction
```

### **#Survival Endpoint**

### **#O' Brien-Fleming efficacy boundaries**

### **# Conditional power method**

```
cp=function(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
w=0.70711,cP=0.9){
```

```
set.seed(1736)
```

```
FSP=0; ESP=0; Nmean=0; power=0
```

```
m.strategy=(k*ma.pos+(1-k)*mb.neg)
m.reverse=(k*mb.pos+(1-k)*ma.neg)
```

```
nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)
```

```
hr=m.strategy/m.reverse
```

```
prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-
tAcr))-exp((-log(2)/m.strategy)*tStd)))
```

```
prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-tAcr))-
exp((-log(2)/m.reverse))*tStd)))
```

```
Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)
```

```
n1=Nclassic*f
```

```
nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients
```

```
rx=log(2)/m.strategy
```

```
ry=log(2)/m.reverse
```

```
diff=rx-ry
```

```
if (case=="null"){rx=ry}
```

```
r=(rx+ry)/2
```

```
expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)
```

```
sigma=r*(1+expTerm)^(-0.5)
```

```
for(iSim in 1:nsim){
```

```
  rx1=rnorm(1,rx,sigma/sqrt(n1))
```

```
  ry1=rnorm(1,ry,sigma/sqrt(n1))
```

```
  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
```

```
  p1=1-pnorm(t1)
```

```
  if (p1>alpha0){
```

```
    FSP=FSP+1/nsim
```

```
    n2=0
```

```
  }
```

```

if (p1<=alpha1){

  power=power+1/nsim
  ESP=ESP+1/nsim
  n2=0
}

if (p1>alpha1 & p1<=alpha0) {

  eSize=diff/sigma

  Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

  n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

  rx2=rnorm(1,rx,sigma/sqrt(n2))
  ry2=rnorm(1,ry,sigma/sqrt(n2))

  t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

  z2=(t1+t2)/sqrt(2);

  p2=1-pnorm(z2);

  if (p2<=alpha2) {power=power+1/nsim}

}

Nmean=Nmean+(n1+n2)/nsim

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.024,w=(1/sqrt(2)),cP=0.8)

```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,  
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

### **#Survival Endpoint**

### **#O' Brien-Fleming efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,  
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,  
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,  
w=0.70711,cP=0.9){
```

```
set.seed(1736)
```

```
FSP=0; ESP=0; Nmean=0; power=0
```

```
m.strategy=(k*ma.pos+(1-k)*mb.neg)  
m.reverse=(k*mb.pos+(1-k)*ma.neg)
```

```
nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)
```

```
hr=m.strategy/m.reverse
```

```
prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-  
tAcr))-exp((-log(2)/m.strategy)*tStd)))
```

```
prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-  
tAcr))-exp((-log(2)/m.reverse)*tStd)))
```

```
Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)
```

```
n1=Nclassic*f
```

```
nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients
```

```
rx=log(2)/m.strategy
```

```
ry=log(2)/m.reverse
```

```
diff=rx-ry
```

```
if (case=="null"){rx=ry}
```

```
r=(rx+ry)/2
```

```
expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)
```

```
sigma=r*(1+expTerm)^(-0.5)
```

```
for(iSim in 1:nsim){
```

```
  rx1=rnorm(1,rx,sigma/sqrt(n1))
```

```
  ry1=rnorm(1,ry,sigma/sqrt(n1))
```

```
  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;
```

```
  p1=1-pnorm(t1)
```



```

if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

if (p1>alpha1 & p1<=alpha0) {

  if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

  er=diff/(abs(rx1-ry1)+0.0000001)

  nfinal=min(nmax, max(Nclassic, (er^2)*Nclassic))

  if (nfinal>n1) {

    rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

    ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

    t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

    z2=(t1+t2)/sqrt(2)

    p2=1-pnorm(z2)

    if (p2<=alpha2) {power=power+1/nsim}

  }

}

Nmean=Nmean+nfinal/nsim

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping

```

```

es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

```

#### Similar codes for 50% and 75%

```

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

```

```

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

```

#### Similar codes for 50% and 75% Information fraction

```

## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3

```

```

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)

```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.002,alpha0=0.124,alpha2=0.0240,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

## **#Binary Endpoint**

### **#Pocock efficacy boundaries**

### **# Conditional power method**

```
cp=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=78,diff=0.15,alpha1=0.0026,alpha0=1,alpha2=0.0240,
  w=1/sqrt(2),cP=0.8){
```

```

set.seed(1736)

FSP=0; ESP=0; nmean=0; power=0

nclassic=((qnorm(1-alpha)+ qnorm(1-beta))^2*(rx_0*(1-rx_0)+ry_0*(1-
ry_0)))/((diff)^2)

nmax=nclassic+100

r=(rx+ry)/2

sigma=sqrt(r*(1-r))

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){

    FSP=FSP+1/nsim
    n2=0
  }

  if (p1<=alpha1){

    power=power+1/nsim
    ESP=ESP+1/nsim
    n2=0
  }
}

```

```

if (p1>alpha1 & p1<=alpha0) {

  eSize=diff/sigma

  Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

  n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

  rx2=rnorm(1,rx,sigma/sqrt(n2))
  ry2=rnorm(1,ry,sigma/sqrt(n2))

  t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

  z2=(t1+t2)/sqrt(2);

  p2=1-pnorm(z2);

  if (p2<=alpha2) {power=power+1/nsim}

}

nmean=nmean+(n1+n2)/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.57,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=0.124,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.72,ry_0=0.57,rx=0.72,
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=0.124,alpha2=0.009,
  w=1/sqrt(2),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

```

```
## For rx=0.74 and ry=0.57
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.57,  
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=0.124,alpha2=0.009,  
  w=1/sqrt(2),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping  
cp(nsim=1000000,alpha=0.025, beta=0.2, ni=0,rx_0=0.74,ry_0=0.57,rx=0.74,  
  ry=0.57,n1=39,diff=0.15,alpha1=0.016,alpha0=0.124,alpha2=0.009,  
  w=1/sqrt(2),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

### **#Binary Endpoint**

### **#Pocock efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,  
  ry=0.57,n1=78, n0=156, diff=0.15,alpha1=0.0026, alpha0=1,  
  alpha2=0.0240){
```

```
  set.seed(1736)
```

```
  FSP=0; ESP=0; nmean=0; power=0
```

```

nclassic=((qnorm(1-alpha)+ qnorm(1-beta))^2)*(rx_0*(1-rx_0)+ry_0*(1-
ry_0))/((diff)^2)

nmax=nclassic+100

r=(rx+ry)/2

sigma=sqrt(r*(1-r))

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(n0, (er^2)*n0))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

      ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

      t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

      z2=(t1+t2)/sqrt(2)

      p2=1-pnorm(z2)

```

```

    if (p2<=alpha2) {power=power+1/nsim}

  }

}

nmean=nmean+nfinal/nsim

}

return(cbind(nclassic,nmean,power,FSP,ESP))

}

## For rx=0.72 and ry=0.57

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.57,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.016, alpha0=0.124,
    alpha2=0.009)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,alpha=0.025, beta=0.2, ni=0, rx_0=0.72,ry_0=0.57,rx=0.72,
    ry=0.57,n1=39, n0=156, diff=0.15,alpha1=0.016, alpha0=0.124,
    alpha2=0.009)

#### Similar codes for 505 and 75% Information fraction

```

### **#Survival Endpoint**

### **#Pocock efficacy boundaries**

### **# Conditional power method**

```

cp=function(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
    w=0.70711,cP=0.9){

```



```

set.seed(1736)

FSP=0; ESP=0; Nmean=0; power=0

m.strategy=(k*ma.pos+(1-k)*mb.neg)
m.reverse=(k*mb.pos+(1-k)*ma.neg)

nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)

hr=m.strategy/m.reverse

prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-
tAcr))-exp((-log(2)/m.strategy)*tStd)))

prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-tAcr))-
exp((-log(2)/m.reverse)*tStd)))

Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)

n1=Nclassic*f

nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients

rx=log(2)/m.strategy
ry=log(2)/m.reverse

diff=rx-ry

if (case=="null"){rx=ry}

r=(rx+ry)/2

expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)

sigma=r*(1+expTerm)^(-0.5)

```

```

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){

    FSP=FSP+1/nsim
    n2=0
  }

  if (p1<=alpha1){

    power=power+1/nsim
    ESP=ESP+1/nsim
    n2=0
  }

  if (p1>alpha1 & p1<=alpha0) {

    eSize=diff/sigma

    Cfun=(qnorm(1-alpha2)-w*qnorm(1-p1))/w

    n2=min(nmax-n1,2*((Cfun-qnorm(1-cP))/eSize)^2);

    rx2=rnorm(1,rx,sigma/sqrt(n2))
    ry2=rnorm(1,ry,sigma/sqrt(n2))

    t2=(rx2-ry2+ni)*sqrt(n2/2)/sigma

    z2=(t1+t2)/sqrt(2);

    p2=1-pnorm(z2);

```

```

    if (p2<=alpha2) {power=power+1/nsim}

  }

  Nmean=Nmean+(n1+n2)/nsim

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3

```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
cp(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

### **#Survival Endpoint**

### **#Pocock efficacy boundaries**

### **# Effect size ratio method**

```
es=function(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=12,tStd=36,ma.pos=6,mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.5,alpha1=0.0026,alpha0=1,alpha2=0.0240,
```

```

w=0.70711,cP=0.9){

set.seed(1736)

FSP=0; ESP=0; Nmean=0; power=0

m.strategy=(k*ma.pos+(1-k)*mb.neg)
m.reverse=(k*mb.pos+(1-k)*ma.neg)

nclassic=2*((qnorm(1-alpha)+ qnorm(1-beta))^2)/((log(m.strategy/m.reverse))^2)

hr=m.strategy/m.reverse

prob.strategy=1-((1/((log(2)/m.strategy)*tAcr))*(exp((-log(2)/m.strategy))*(tStd-
tAcr))-exp((-log(2)/m.strategy)*tStd)))

prob.reverse=1-((1/((log(2)/m.reverse)*tAcr))*(exp((-log(2)/m.reverse))*(tStd-
tAcr))-exp((-log(2)/m.reverse)*tStd)))

Nclassic=nclassic/(0.5*prob.strategy+0.5*prob.reverse)

n1=Nclassic*f

nmax=Nclassic+100 ##### Allow recruitment of additional 100 patients

rx=log(2)/m.strategy
ry=log(2)/m.reverse

diff=rx-ry

if (case=="null"){rx=ry}

r=(rx+ry)/2

expTerm=exp(-r*tStd)*(1-exp(r*tAcr))/(tAcr*r)

sigma=r*(1+expTerm)^(-0.5)

```

```

for(iSim in 1:nsim){

  rx1=rnorm(1,rx,sigma/sqrt(n1))

  ry1=rnorm(1,ry,sigma/sqrt(n1))

  t1=(rx1-ry1+ni)*sqrt(n1/2)/sigma;

  p1=1-pnorm(t1)

  if (p1>alpha0){FSP=FSP+1/nsim; nfinal=n1}

  if (p1<=alpha1){power=power+1/nsim; ESP=ESP+1/nsim; nfinal=n1}

  if (p1>alpha1 & p1<=alpha0) {

    if (diff*(rx1-ry1+ni)<0) {nfinal=n1}

    er=diff/(abs(rx1-ry1)+0.0000001)

    nfinal=min(nmax, max(Nclassic, (er^2)*Nclassic))

    if (nfinal>n1) {

      rx2=rnorm(1,rx,sigma/sqrt(nfinal-n1))

      ry2=rnorm(1,ry,sigma/sqrt(nfinal-n1))

      t2=(rx2-ry2+ni)*sqrt((nfinal-n1)/2)/sigma

      z2=(t1+t2)/sqrt(2)

      p2=1-pnorm(z2)

      if (p2<=alpha2) {power=power+1/nsim}

    }

  }

  Nmean=Nmean+nfinal/nsim

```

```

}

return(cbind(nclassic,Nclassic,hr,Nmean,power,FSP,ESP))

}

## For ma.pos=6, ma.neg=6.7, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6,
mb.neg=4.3, mb.pos=4.3, ma.neg=6.7,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.7, ma.neg=6, mb.neg=4.3, mb.pos=4.3

#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.7,
mb.neg=4.3, mb.pos=4.3, ma.neg=6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)

#### Similar codes for 50% and 75% Information fraction

## For ma.pos=6.6, ma.neg=6.3, mb.neg=4.3, mb.pos=4.3

```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.6,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.3,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```

```
## For ma.pos=6.3, ma.neg=6.6, mb.neg=4.3, mb.pos=4.3
```

```
#### Test null hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="null",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Test alternative hypothesis__Information fraction=25%__No futility stopping
es(nsim=1000000,case="alternative",alpha=0.025, beta=0.2,
ni=0,k=0.2,tAcr=104.28,tStd=116.28,ma.pos=6.3,
  mb.neg=4.3, mb.pos=4.3, ma.neg=6.6,
f=0.25,alpha1=0.016,alpha0=0.124,alpha2=0.009,w=(1/sqrt(2)),cP=0.8)
```

```
#### Similar codes for 50% and 75% Information fraction
```



### D.3.3. Type I error probabilities

---

The type I error probabilities derived from O'Brien-Fleming and Pocock type spending functions for a two-stage design with one-sided 0.025 significance level can be calculated in R software by the following codes:

```
2-2*pnorm(qnorm(1-0.025/2)/sqrt(1/2)) # O'Brien-Fleming type boundary in stage 1
```

```
0.025-(2-2*pnorm(qnorm(1-0.025/2)/sqrt(1/2))) # O'Brien-Fleming boundary in stage  
#2
```

```
0.025*log(1+((exp(1)-1)*(1/2))) # Pocock boundary in stage in stage 1
```

```
0.025-(0.025*log(1+((exp(1)-1)*(1/2)))) # Pocock boundary in stage 2
```

In case of two-sided 0.05 significance level, the aforementioned one-sided stopping boundaries would be multiplied by 2. Additionally, the aforementioned calculations can be performed by the package GroupSeq in R statistical software which computes probabilities regarding group sequential designs including the O'Brien-Fleming and Pocock type spending functions.